GENOMIC AND TRANSCRIPTOMIC CHARACTERIZATION OF INFLAMMATORY BOWEL DISEASE

A Dissertation
Presented to
The Academic Faculty

by

Angela Mo

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in Bioinformatics in the
School of Biological Sciences

Georgia Institute of Technology
May 2021

GENOMIC AND TRANSCRIPTOMIC CHARACTERIZATION OF INFLAMMATORY BOWEL DISEASE

Approved by:

Dr. Gregory Gibson, Advisor
School of Biological Sciences
Georgia Institute of Technology

Dr. Peng Qiu
Department of Biomedical Engineering
Georgia Institute of Technology

Dr. I. King Jordan
School of Biological Sciences
Georgia Institute of Technology

Dr. Subra Kugathasan
Department of Pediatrics and Human Genetics
Emory University

Dr. Joseph Lachance
School of Biological Sciences
Georgia Institute of Technology

Date Approved: April 2, 2021

To my family and friends

ACKNOWLEDGEMENTS

When I first set foot on Georgia Tech's campus in 2013, I had little idea of the seven-year journey ahead of me. I am grateful to my then-professor, now-advisor Dr. Greg Gibson for first kindling my fascination with genomics as a young undergraduate student, and for his subsequent guidance and support through my development into a full-fledged scientist.

I would like to thank the members of my committee, Dr. Subra Kugathasan, Dr. King Jordan, Dr. Joe Lachance, and Dr. Peng Qiu, for their wisdom and advice imparted over the years.

I feel incredibly fortunate to have been surrounded by my colleagues in the lab, who are not only passionate scientists but also the most wonderfully supportive friends. I especially want to thank Urko, Ruoyu, Sini, and Maggie, the residents of 2202 who have always filled the room with laughter and thoughtful discussion.

Finally, I would like to thank my family for their constant support, and always encouraging me to strive to be the best person I can be.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS AND ABBREVIATIONS

AA	African American
ANOVA	Analysis of Variance
BIT	Blood Informative Transcript
BTM	Blood Transcription Module
CD	Crohn's Disease
CEU	Northern European from Utah
CPM	Counts per Million
eGene	Expression Gene
eQTL	Expression Quantitative Trait Locus
FDR	False Discovery Rate
GATK	Genome Analysis Toolkit
GO	Gene Ontology
GrCh38	Human Genome Assembly 38
GSEA	Gene Set Enrichment Analysis
GTEx	Genotype-Tissue Expression Project

GWAS	Genome-wide Association Study
hg19	Human Genome Assembly 19
HLA	Human Leukocyte Antigen
IBD	Inflammatory Bowel Disease
JIA	Juvenile Idiopathic Arthritis
LD	Linkage Disequilibrium
MAF	Minor Allele Frequency
mQTL	Methylation Quantitative Trait Locus
PBMC	Peripheral Blood Mononuclear Cell
PCA	Principal Component Analysis
RA	Rheumatoid Arthritis
scRNA-seq	Single Cell RNA Sequencing
SNP	Single Nucleotide Polymorphism
TA	Transit-Amplifying
TMM	Trimmed Mean of M-Value
TRS	Transcriptional Risk Score
UC	Ulcerative Colitis

WGS Whole Genome Sequencing

YRI Yoruban African

SUMMARY

Inflammatory bowel disease (IBD) is a chronic idiopathic disorder resulting in the inflammation of the gastrointestinal tract, which encapsulates both Crohn's disease (CD) and ulcerative colitis (UC). There is strong evidence of familial aggregation of IBD, and more than 200 genetic variants have been identified as associated with IBD (1, 2). Currently, more than 1.5 million individuals in the United States suffer from IBD, with prevalence of the disease on the rise, particularly in minority populations (3). Although African American and Caucasian populations in the US share a similar burden of disease, studies suggest that African American CD patients are at a greater risk for disease complications and often experience worse outcomes when compared with Caucasian patients (4-6). Despite this, African Americans are greatly underrepresented in clinical trials and research studies on IBD, where the majority of genetic contributors to disease have been identified in cohorts of exclusively Caucasian individuals of European descent. Resolving this disparity is critical to determining whether there are biological mechanisms underlying CD in African Americans which differ from those in Caucasians.

The primary question driving this thesis is whether genomic and transcriptomic profiling have the potential to direct personalized therapeutic interventions for IBD patients. Prediction of disease course is especially relevant in IBD, because early, appropriate introduction of anti-TNFα therapy can slow the progression of the disease in patients who would otherwise experience severe flares, bowel penetration, or require invasive surgeries. Conversely, early prediction can also identify patients who are not expected to progress and thus help to avoid unnecessary, harmful, and costly therapy with

biologics. Our prior studies have demonstrated that transcriptomic data can be used to identify patients who are likely to progress from B1 stable CD to B3 penetrating CD, and that gene expression can also be linked to GWAS via Transcriptional Risk Scores (TRS) (7, 8).

In this thesis, I first describe a comparative analysis of gene expression and eQTL in IBD and juvenile idiopathic arthritis, another clinically heterogeneous immune-related disorder. Next, I examine the influence of African ancestry proportion on gene expression in the ileum of Crohn's disease patients. I then discuss the utility of gene expression at time of disease diagnosis to predict risk of progression to colectomy in an inception cohort of pediatric ulcerative colitis patients. Finally, I present an exploration of cell type composition of ileal epithelial cells at single-cell resolution in healthy, treatment-naïve, and treated Crohn's disease. I conclude with a summary of the progress achieved and challenges to be overcome for the successful application of genomics for precision medicine in IBD and beyond.

CHAPTER 1. INTRODUCTION

Inflammatory bowel diseases, comprised mainly of Crohn's disease and ulcerative colitis, are chronic relapsing and remitting diseases of unknown etiology which result in the inflammation of the gastrointestinal tract. Historically, IBD was considered to be a disease of westernized nations, but recent epidemiological studies have demonstrated the growing impact of IBD mirroring trends in global socioeconomic development (9). In the United States, it is now estimated that more than 1.5 million individuals suffer from IBD, with prevalence of the disease continuing to rise (3). Our healthcare system has been tasked with the challenge of providing quality, cost-effective, long-term care, which begins with

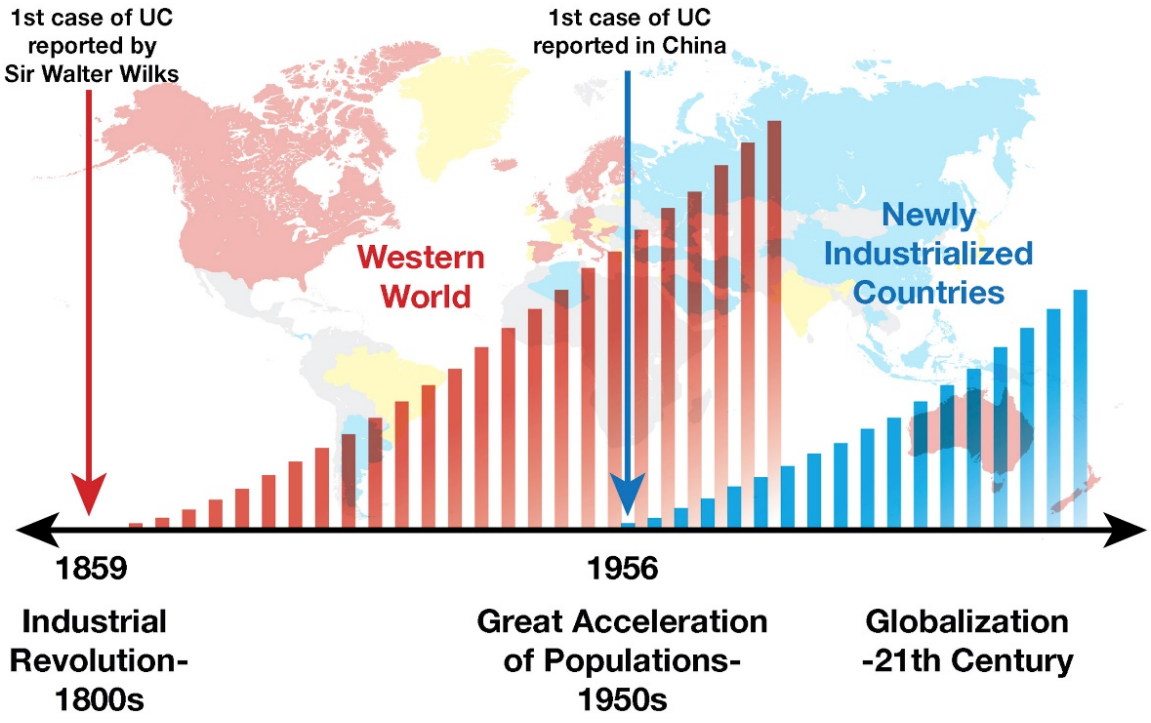


Figure 1 – The growing global impact of IBD. Incidence and prevalence of IBD increasing in newly industrialized countries echoes historical trends of industrialized countries. From Kaplan et al., (9).

understanding the underlying biological mechanisms of this complex disease. To that end, validation of the utility of transcriptomic profiling at diagnosis for prediction of disease course, evaluating the influence of African ancestry on disease via transcriptomic profiling, and identifying the individual contributions of immune cell types to transcriptomic signatures of disease progression will offer critical insights.

1.1 Influencing the Natural History of Inflammatory Bowel Disease

The two most common forms of IBD are Crohn's disease and ulcerative colitis. Inflammation in UC is typically restricted to the mucosal and submucosal layer of the colon, while CD may affect the entirety of the gastrointestinal tract (10, 11). Physicians attempt to classify subtypes of disease with clinical indices and administer standardized therapeutic regimens accordingly.

1.1.1 Clinical classifications of IBD

Current indices of IBD severity such as PUCAI (Pediatric Ulcerative Colitis Activity Index) and CDAI (Crohn's Disease Activity Index), are typically based on clinical phenotypes and markers like rectal bleeding, albumin, and C-reactive protein levels (12, 13). A scoring system known as the Lémann Index was also recently developed to provide multiple assessments of bowel damage over the course of CD (14). One of the measures of activity included within the Lémann Index is disease behavior, where non-stricturing and non-penetrating disease is termed "B1", stricturing disease is termed "B2", and penetrating

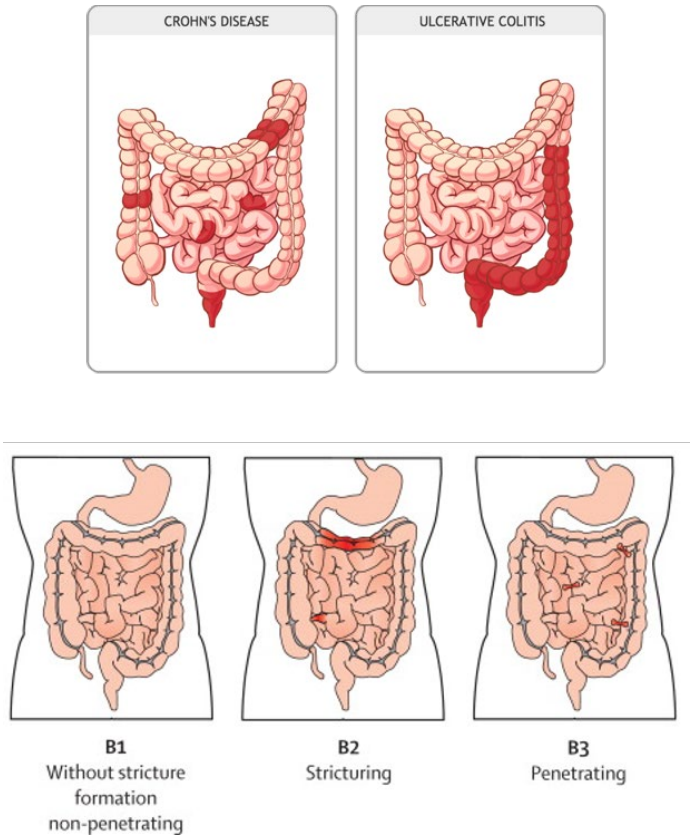


Figure 2 – Subtypes of IBD. Top row: Crohn's disease presents patchy inflammation throughout the GI tract, while UC is typically restricted to the colon. Illustration by The Hospital for Sick Children (11). Bottom row: Lémann index classifications. From Baumgart et al. (15).

disease is termed "B3", in order of least to greatest severity (15). At diagnosis, CD patients typically present with inflammatory-only B1 disease, with an estimated 50% of patients proceeding to complicated disease within five years (16). As disease progresses to B2 and B3, blockages of the bowel occur which must be addressed with surgical resection in up to 80% of patients with complicated disease (17).

1.1.2 Standard therapeutics for IBD

The goal of therapy for CD is to maintain disease at the stable B1 state and delay or eliminate the need for invasive surgical procedures. Some of the commonly used medications for treatment of IBD include aminosalicylates, corticosteroids, immunomodulators like methotrexate and azathioprine, and biologics such as infliximab and vedolizumab. Of particular interest in current pharmacological research are anti-TNFα therapies. Tumor necrosis factor-α plays a key role in mucosal inflammation, and is thought to mediate the inflammatory cascade in IBD by inducing the activation of various immune cells, including neutrophils, monocytes, macrophages, and lymphocytes (18). Infliximab, a common anti-TNFα drug used in IBD, binds to TNF-α, preventing its binding with receptors and thereby inducing the apoptosis of the immune cells causing inflammation. Prior studies have shown that early treatment with anti-TNFα therapy is associated with a reduction in disease complications later on (7, 19-21). However, aggressive drug therapy of all patients is not a feasible solution due to toxicity and economic burden. The need for tools capable of early discrimination of patients with poor prognosis to direct prompt therapeutic decision making is clear.

1.2 Tools for Studying the Genetics of Disease

Three major approaches capture complementary information key to characterizing the genetics of IBD—bulk transcriptomic profiling, single cell transcriptomic profiling, and expression quantitative trait loci (eQTL) analysis.

1.2.1 Bulk RNA-sequencing

Bulk transcriptomic profiling uses RNA-seq to quantify the levels of gene expression present in a biological sample. Classically, the central dogma of biology states that information stored in genes flows from DNA to RNA, and is then transcribed into proteins (22). Messenger RNA (mRNA) are the molecules complementary to DNA which serve as the intermediate between DNA and protein and collectively constitute the transcriptome. Historically, methods like hybridization-based microarrays were utilized to perform gene expression profiling, but had drawbacks such as high technical variability and limitations in probe design (23). Next-generation sequencing is the successor to those technologies and offers considerable advantages over prior techniques. RNA-seq starts with the isolation of RNA from a tissue sample. An RNA-seq library is generated by purifying down to the desired RNA molecules, typically mRNA, then reverse-transcribing the RNA into cDNA. Tens of millions of RNA-seq reads per sample are aligned to a reference genome, then transcript length and GC content biases can be adjusted for computationally in subsequent normalization procedures (24, 25). Commonly, differential gene expression analysis then follows to identify genes which are statistically distinguished between groups of interest.

1.2.2 Single-cell RNA-sequencing

One caveat of bulk transcriptomic profiling is that gene expression varies greatly between different tissue and cell types, which can lead to loss of detection of genuine biological signatures. Compared to bulk RNA-seq, the relatively novel technique of single

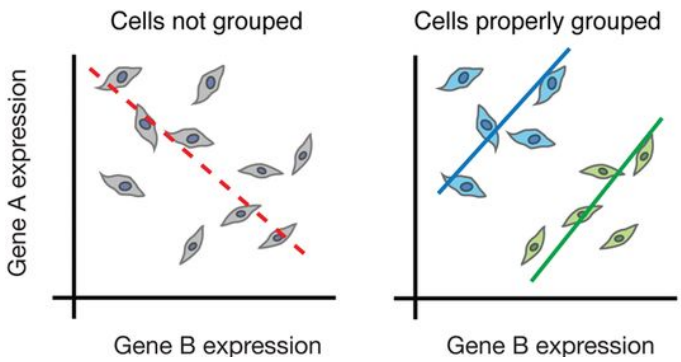


Figure 3 – The potentially signal-obscuring effects of grouped gene expression measurements. From Trapnell (26).

cell RNA-seq offers much higher resolution (26). The basic principles of single cell RNA-seq are similar, albeit with two additional challenges—isolating single cells and amplifying the much smaller amounts of mRNA (27). Appropriately adjusted single-cell data has the potential to reveal rare cell populations and track the developmental trajectories of certain cell lineages (28). However, normalization and analysis methods are not as well established as in bulk RNA-seq studies. Due to the sparsity of single cell RNA-seq data, normalization methods previously developed for use in bulk RNA-seq studies can be inaccurate when applied to single-cell data (29). Numerous single cell specific bioinformatics methods are currently being developed; two examples of such tools include Seurat and Monocle (30, 31).

1.2.3 Mapping of eQTL

Combining genotype data with gene expression data also enables the discovery of genetic loci associated with differences in gene expression, or expression quantitative trait loci (eQTL). Traditionally, eQTL studies have been performed with bulk RNA-seq data,

which has uncovered numerous associations. For example, the Blood eQTL Browser is based on a genome-wide analysis which identified cis-eQTL for 44% of tested genes (32). Another large consortium, the Genotype-Tissue Expression (GTEx) Project, aims to characterize genetic effects on gene expression in a tissue-specific manner, and has catalogued effects across 44 human tissues (33). It is also possible to perform cell type specific eQTL studies. Previously, single-cell eQTL analyses were performed by deconvoluting bulk gene expression data, or by using purified cells from sorting techniques like FACS, as was done for the Database of Immune Cell Expression (DICE) project. In their first report, the DICE group examined 13 immune cell types and 2 activated cell types, and identified cis-eQTL for 12,154 genes, 41% of which were specific to a unique cell type (34). However, these methods can be limited by dependence on surface marker genes and exclusion of rare cell populations (35). Utilizing single cell RNA-seq theoretically enables the discovery of eQTL associations without these biases but is more complicated in practice. A proof of concept study in 2018 used single cell RNA-seq data of 25,000 PBMCs from 45 individuals to validate associations identified previously in whole blood eQTL studies, but faced difficulties when investigating cell type dependent associations because of the high levels of dropout characteristic of single cell RNA-seq data (36). More recently, the single-cell eQTLGen Consortium was established to systematically analyze single cell expression of peripheral blood mononuclear cells, building upon knowledge previously collected in the bulk whole blood eQTLgen Consortium (37). The OneK1K study also represents another large-scale study of over 1.2 million PBMCs from nearly one thousand individuals aiming to characterize cell type specific influences of genetic loci on gene

expression (38). As the cost of single-cell sequencing technologies drops and new techniques are developed to address current challenges in analysis and replicability, similar groups will continue to emerge to synthesize single-cell RNA-seq datasets across additional tissues and fully characterize gene regulation at the single cell level.

1.3 Successive Efforts to Leverage Genetics in the Study of IBD

As the cost of nucleic acid sequencing has fallen, the potential for genomic and transcriptomic profiling to direct and improve clinical care has risen. In the field of inflammatory bowel disease, ever-evolving approaches to characterizing the genetic architecture underlying disease risk and pathogenesis continue to be successfully applied to larger and larger cohorts of patients.

1.3.1 GWAS identify IBD risk loci

Following the sequencing of the first human genome in the early 2000s, rapid advancements in sequencing technologies enabled novel mapping of genotype to phenotype associations via genome-wide association studies (GWAS). GWAS have been particularly successful within the field of inflammatory bowel disease. The power of GWAS to identify risk loci has increased with both technological improvements in genotyping arrays and increasing numbers of IBD patients. Numerous GWAS on Crohn's disease, ulcerative colitis, and combined inflammatory bowel diseases can be summarized by three landmark studies: Jostins et al., Liu et al., and most recently de Lange et al.,

culminating in a total of approximately 240 variants associated with risk for inflammatory bowel disease identified to date (1, 2, 39). Cumulatively, well over 100,000 individuals have been profiled in these studies. The authors report overlap of susceptibility loci with other inflammatory and immune-mediated diseases, including ankylosing spondylitis and psoriasis. De Lange et al. in particular identified associations located near integrin genes, notable because of the recent emergence of drugs such as vedolizumab that prevent immune cell migration to the gut by blocking integrin receptors (39).

Despite the impressive size and achievements of these recent integrated studies, some of the fundamental weaknesses of GWAS remain to be addressed. First is the issue of missing heritability, the proportion of heritability left unexplained by GWAS which may be attributed to poor detection of rare variants, common variants with low individual effects, and unaccounted environmental or gene interactions (40). In inflammatory bowel disease specifically, estimations of heritability of risk for CD and UC based on GWAS are 37% and 27% respectively, just under half of the heritability calculated from twin studies, 75% and 67% (41, 42). Second, the great majority of GWAS, including GWAS of IBD, have been conducted in cohorts of primarily European and secondarily East Asian populations, limiting applications in populations of other ancestries (43). This bias will be further addressed in the following section on inflammatory bowel disease in African Americans. Finally, one of the major weaknesses of genome-wide association studies is the missing link between identified risk variants and underlying causal mechanisms of disease. Genetic risk scores based exclusively on GWAS significant SNPs have

demonstrated limited predictive power to distinguish disease risk, further highlighting the necessity for complementary information to translate research into reality (44).

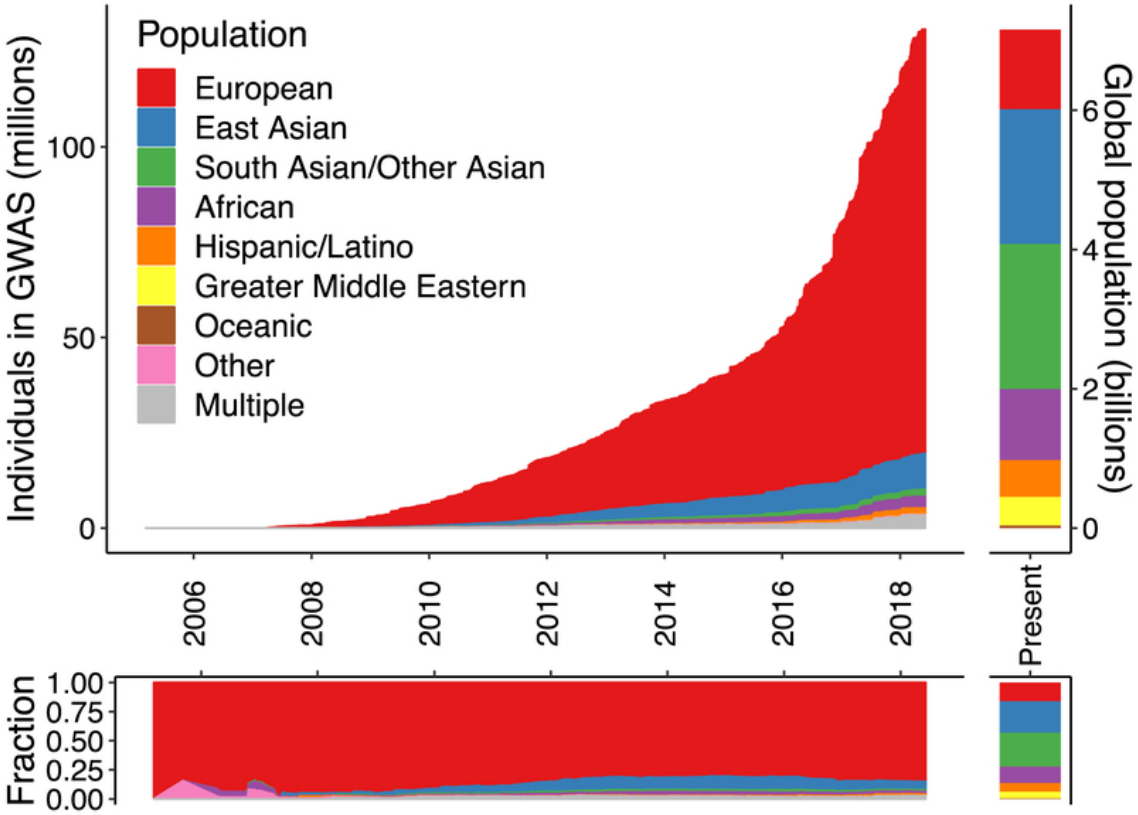


Figure 4 – Population bias in GWAS. GWAS population proportions contrasted against the global population. From Martin et al. (43).

1.3.2 Transcriptomics offers insights into causal mechanisms

Bridging the gap between GWAS and biological interpretation, gene expression and more recently single-cell gene expression studies offer insights into potential mechanisms of disease. Bulk tissue RNA-Seq studies have previously revealed differential expression of key inflammatory and immunomodulatory genes distinguishing subtypes of disease (7, 45-47). Newer single-cell profiling has enabled the dissection of more refined

cell-type specific contributions to disease (48). For example, Parikh et al. sampled colonic tissue from three healthy individuals and inflamed and uninflamed colonic tissue from three individuals with ulcerative colitis (49). They identified two clusters representing inflammation-associated goblet cells and intraepithelial immune cells, and discovered 1,147 differentially expressed genes in inflamed UC samples. Individual cell subtypes such as colonocytes, goblet cells, and Paneth cells each induced different pathways contributing to overall signatures of disease. Additionally, the authors tested UC risk loci identified in GWAS for cell-type expression specificity and found that while intra-epithelial T cells were most associated with IBD in healthy tissue, diverse subsets of immune and absorptive epithelial cell types each contributed small defects in function to the overall failure of the epithelial barrier seen in disease.

Similarly, Smillie et al. sequenced 366,650 cells obtained from a collection of 68 paired colonic biopsies from 18 ulcerative colitis patients and 10 healthy controls (50). They characterized 51 subsets of cells based on gene expression, and identified changes in cell type composition between non-inflamed, inflamed, and healthy samples. Additionally, they studied the cell type specific expression of GWAS-implicated genes and discovered enrichment both in a cell type and disease specific manner. Their findings both confirmed previously reported associations and revealed novel associations that could have only been detected with single cell resolution (51, 52).

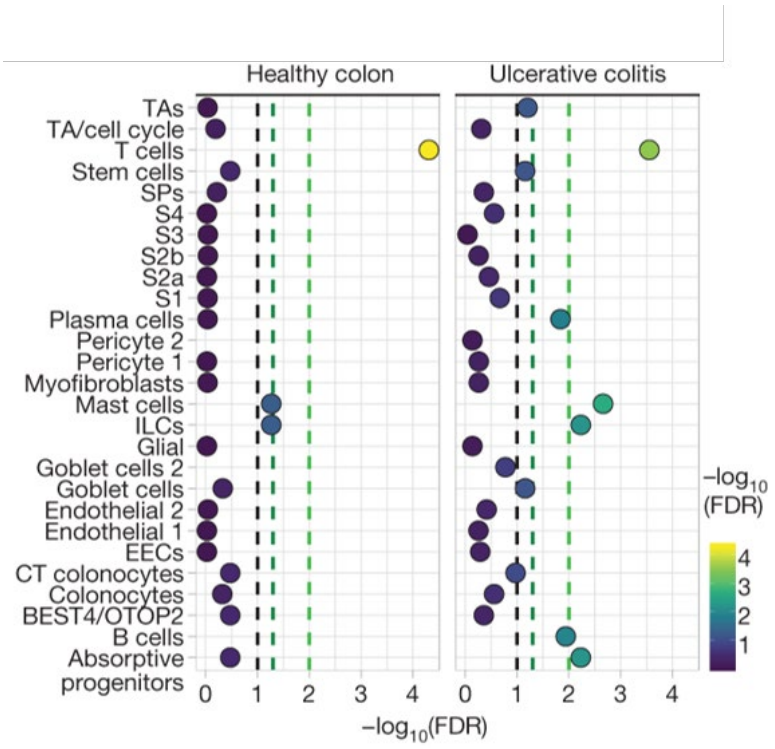


Figure 5 – Tissue-specific enrichment of UC GWAS loci in single-cell clusters in colonic epithelium and mesenchyme in healthy and active UC. From Parikh et al. (49).

1.4 Inflammatory Bowel Disease in African Americans

Approximately 40 million individuals who self-identify as African American live in the United States. The influence of ethnicity on IBD phenotypes remains poorly characterized, as the great majority of studies are based on cohorts of exclusively white individuals, and what research does exist is limited by small sample sizes. Although this discrepancy is not by any means unique to the field of IBD, it presents challenges to understanding the etiology of disease in African Americans, as evidence exists that there may be population-specific differences in risk and outcomes. It has been estimated, fairly

consistently, that the prevalence of IBD is approximately two to three times greater in Caucasians compared with African Americans (53-55). However, the literature offers conflicting reports on prognosis in African Americans. Several studies claim to observe no significant differences in outcomes between Caucasian and African American patients (56). Yet others report that African Americans experience more severe disease, worse outcomes, and require more intensive medical interventions. For example, one study of racial disparities in pediatric CD found that black children had a shorter time to first hospital readmission and greater risk of readmission, with longer lengths of hospitalization, findings seemingly corroborated by other hospital-based studies in adult cohorts (4, 57, 58). In addition, differences in clinical phenotypes such as perianal disease, occurrence of fistulas, and ileal involvement have also been noted (5, 59, 60). African American race was also found to be associated with increased risk of progressing to stricturing or penetrating disease in CD, and increased rate of postoperative complications (7, 60). These findings suggest that the biological mechanisms of disease in IBD may differ between populations, but purely epidemiological and clinical data is insufficient to draw conclusions. It is necessary to examine the underlying genetics of IBD to separate the contribution of factors like socioeconomic status and healthcare utilization from intrinsic ancestry-specific differences in disease.

Immune response is a highly complex phenotype that has been shaped by natural selection over the course of many generations. Conducting genetic research in African Americans requires consideration of the unique features and evolution of the population. The genomes of African Americans are admixed, composed of approximately 80% West

African and 20% European ancestry (61). Owing to their high proportion of West African ancestry, African American genomes are much more diverse than Caucasian genomes, and possess shorter linkage disequilibrium (LD) blocks (62). Studies of immune response variation between populations have consistently shown that individuals of African descent have stronger responses to bacterial and viral challenges (63, 64). Underlying this ancestry-specific variation are antiviral and inflammatory-related genes enriched for eQTL associations, and changes in allele frequencies (64). Increased diversity within the major histocompatibility complex (MHC), a critical region of the genome that has been associated with numerous autoimmune diseases, in African American populations is well documented and may also account for variations in immune response (65-68).

Some of the known risk factors which contribute to the risk of disease progression in IBD include age at diagnosis, inflammation location, response to microbial antigens, and variants associated with certain genes such as *NOD2* and *MMP3* (69-72). Current research has identified variants associated with IBD risk that appear to be common amongst African American and Caucasian populations, as well as ancestry specific variants. *NOD2*, nucleotide oligomerization domain, is the gene most consistently associated with risk for complications in white and especially Jewish populations (73). Multiple studies have concluded that *NOD2* mutations are solely due to European admixture, the variants confer a similar increase in risk compared with whites, but account for a lower attributable amount of risk due to reduced allele frequencies in African American populations (74-76). Numerous large genome-wide association studies have been performed in IBD in mostly European populations, yielding 241 susceptibility loci from a total sample size of nearly

60,000 individuals (39, 77). An initial GWAS of IBD in African Americans with a total sample size of approximately 3,000 individuals replicated several European loci but was underpowered to detect novel associations; however, a newer study expanded to more than 7,000 individuals identified two novel African-specific UC loci associated with *ZNF649* and *LSAMP*, and multiple associations in the HLA region for IBD which were African-specific (78, 79). A very recent whole-genome sequencing study of over 3,000 individuals established genome-wide significant association of *PTGER4* with Crohn's disease in African Americans for the first time (80). Importantly, the authors demonstrated that utilizing ancestry-matched weights for the generation of polygenic risk scores significantly improved performance in the highest risk percentiles. Further investigation into the function of risk variants is necessary to understand their contributions to disease.

1.5 Personalized Medicine for IBD with Genomic & Transcriptomic Profiling

As mentioned previously, early therapy with anti-TNFα can influence the course of CD and reduce the risk of developing complications. A recent study found that the per-member per-year cost of the average IBD patient taking biologics in 2015 was $36,051, a nearly 8-fold increase in costs compared with patients taking 5-ASA only and 36-fold increase compared with patients taking immunomodulators only (81). In addition, the potential off-target effects of anti-TNFα include immunogenicity, increased rate of infections, and organ failure, making blanket drug prescription unconscionable (82). The limitations of anti-TNFα therapy highlight the need for the development of tools for the

early identification of patients at risk for progressing to complicated disease and thus who would benefit the most from treatment with biologics.

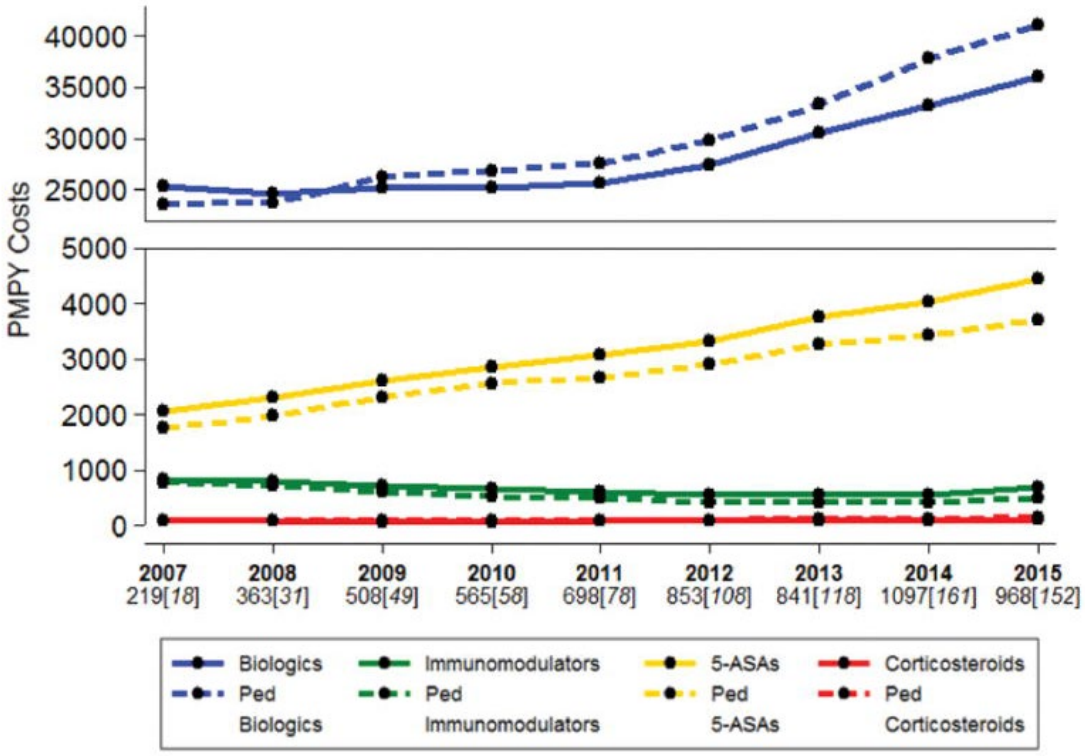


Figure 6 – Per-member per-year costs of common IBD medications. The cost of treatment with biologics far outpaces the cost of other medications. From Yu et al (81).

Although GWAS has successfully identified thousands of variants associated with disease, the predictive capability of genetic risk scores is hampered by the low percentage of heritability explained (40). Gene expression signatures are often associated with outcomes of interest, as is the case in progression of disease in Crohn's, but it is difficult to ascertain causality of specific genes. By integrating genomic and transcriptomic data through eQTL studies and beyond, we may be able to capture the full potential of both to further precision medicine for IBD (83). For example, Transcriptional Risk Score (TRS) is

an approach which successfully combines GWAS results, eQTL information, and gene expression to identify patients who will progress to complicated CD, effectively linking statistical associations to underlying biological mechanisms (8). Holistically, this thesis seeks to achieve this type of integration of genetics and the transcriptome for characterization of disease, with additional foci on elucidating the influences of ancestry and cell specificity.

CHAPTER 2. DISEASE-SPECIFIC REGULATION OF GENE EXPRESSION IN A COMPARATIVE ANALYSIS OF JUVENILE IDIOPATHIC ARTHRITIS AND INFLAMMATORY BOWEL DISEASE

The genetic and immunological factors that contribute to differences in susceptibility and progression between sub-types of inflammatory and autoimmune diseases continue to be elucidated. Inflammatory bowel disease and juvenile idiopathic arthritis are both clinically heterogeneous and known to be due in part to abnormal regulation of gene activity in diverse immune cell types. Comparative genomic analysis of these conditions is expected to reveal differences in underlying genetic mechanisms of disease.

We performed RNA-Seq on whole blood samples from 202 patients with oligoarticular, polyarticular, or systemic juvenile idiopathic arthritis, or with Crohn's disease or ulcerative colitis, as well as healthy controls, to characterize differences in gene expression. Gene ontology analysis combined with Blood Transcript Module and Blood Informative Transcript analysis was used to infer immunological differences. Comparative expression quantitative trait locus (eQTL) analysis was used to quantify disease-specific regulation of transcript abundance.

A pattern of differentially expressed genes and pathways reveals a gradient of disease spanning from healthy controls to oligoarticular, polyarticular, and systemic juvenile idiopathic arthritis (JIA); Crohn's disease; and ulcerative colitis. Transcriptional risk scores also provide good discrimination of controls, JIA, and IBD. Most eQTL are

found to have similar effects across disease sub-types, but we also identify disease-specific eQTL at loci associated with disease by GWAS. JIA and IBD are characterized by divergent peripheral blood transcriptomes, the genetic regulation of which displays limited disease specificity, implying that disease-specific genetic influences are largely independent of, or downstream of, cis-eQTL effects.

This study was performed in collaboration with the Prahalad lab and Kugathasan lab at Emory University. Our findings were published in *Genome Medicine* (84).

2.1 Introduction

While genomic analyses have clearly established a high degree of shared genetic susceptibility across autoimmune and inflammatory disorders, the reasons for disease-specific effects of particular loci are yet to be understood (85). Likely explanations range from the technical, such as variable statistical power across studies, to the biological, including restriction of effects to relevant cell types for each condition, and interactions between genotypes and either the environment or genetic background. Since the majority of genome-wide association study (GWAS) associations are likely regulatory, attention has focused on mapping genetic effects on gene expression and/or epigenetic marks, namely discovery of expression quantitative trait locus (eQTL) and their methylation counterparts, mQTL (86). With a few exceptions, most studies attempting to relate GWAS to functional genomics have utilized large public eQTL and epigenetic datasets of peripheral blood-derived profiles of healthy volunteers. These implicitly assume equivalence of eQTL across health and disease, despite recent findings that eQTL can be modified by ex vivo

treatments which mimic perturbations corresponding to disease states (63, 87). In order to evaluate the ratio of common to disease-specific effects in inflammatory autoimmune disease, here we describe side-by-side comparative eQTL analysis of juvenile idiopathic arthritis (JIA) and inflammatory bowel disease (IBD), also comparing the transcriptomes among major sub-types within both JIA and IBD.

IBD has been extensively studied using a variety of genomic approaches, but despite several early publications, JIA has been less well characterized (88-91). JIA is the most common rheumatic disease of childhood, with an estimated prevalence of approximately 1.2 individuals per 1000 in the USA (92). It comprises multiple clinically and genetically distinct forms of arthritis with onset prior to age 16. Although all forms of JIA are characterized by persistent swelling of the joints, the disease is further classified into sub-types based on clinical presentation (93). Oligoarticular JIA affects four or fewer joints and is the most common and typically the mildest form of JIA (93, 94). Polyarticular JIA involves five or more joints and is intermediate in severity. Both oligoarticular and polyarticular JIA disproportionately affect females. Systemic JIA (sJIA) is distinct from other JIA sub-types, displaying unique symptoms and no bias towards females (93, 95). Diagnosis is based on presentation of arthritis accompanied by spiking fever, rash, and lymphadenopathy. Approximately 10% of sJIA patients are also diagnosed with life-threatening macrophage activation syndrome, and about 50% experience a persistent course of disease and are unable to achieve remission (95, 96).

The categorization of sub-types based primarily on clinical criteria reflects uncertainty about the biological factors that contribute to the heterogeneity of the disease.

The immune system is thought to play a critical role in the pathogenesis of JIA. Levels of immune-related cells like lymphocytes, monocytes, and neutrophils are differentially elevated between sub-types (97), as is also seen in other autoimmune and autoinflammatory diseases such as rheumatoid arthritis (RA) and inflammatory bowel disease (98). Evidence of T cell activation has been described in oligoarticular and polyarticular patients, suggesting the importance of adaptive immunity in these sub-types (94, 99), but there is considerable heterogeneity in immune profiles that masks differences between levels of severity (100, 101), with age-of-onset also an important factor influencing gene expression (102). In contrast, sJIA is thought to be more characterized by activation of innate immunity and upregulated monocytes, macrophages, and neutrophils (95, 103).

Extensive genome-wide association studies have been performed across autoimmune classes and are conveniently summarized on the ImmunoBase website (https://genetics.opentargets.org/immunobase), which as of February 2018 lists 23 validated loci for JIA, 81 for RA, 102 for ulcerative colitis (UC), and 122 for Crohn's disease (CD) (104). Previous studies have demonstrated familial aggregation of JIA, supporting the idea that genetics plays a role in susceptibility (105) as well as sub-type development. Studies of genetic variants within the major histocompatibility complex region have uncovered associations between various human leukocyte antigen (HLA) polymorphisms and sub-types of JIA (106, 107). HLA-independent loci such as PTPN22 and STAT4 have also been repeatedly found in genome-wide association studies to be associated with oligoarticular and RF-negative polyarticular JIA at genome-wide significance levels (108-111), while polymorphisms in interleukins 1 and 10 were early on

identified as occurring at higher frequencies in sJIA patients (112, 113). The most recent international GWAS of 982 children with sJIA concluded that the systemic form of JIA engages more inflammatory than autoimmune-related genes (114), consistent with clinical observations of the course of disease.

Diverse autoimmune conditions certainly are attributable in part to intrinsic aspects of the focal tissue and in part to gene activity in the immune system, some of which should be detectable in peripheral blood samples. It is thus surprising that side-by-side comparisons of immune gene expression across disease sub-types have not been reported. Transcriptomic studies of disease are for practical reasons orders of magnitude smaller than GWAS, typically involving fewer than 200 patients, but these are nevertheless sufficient to identify eQTL given the relatively large effect of regulatory polymorphisms on local gene expression. Numerous blood- and tissue-specific susceptibility loci and eQTL have previously been discovered (115-117). It is likely that sJIA in particular shares associated risk polymorphisms with IBD given the auto-inflammatory component of both diseases. For instance, a mutation in LACC1 that was initially associated with Crohn's disease was later found also to be associated with sJIA (118, 119). Thus, IBD is an attractive candidate for comparison with JIA to elucidate the mechanisms behind each of the sub-types. Here we contrast healthy controls; patients with oligoarticular, polyarticular, or systemic JIA; and patients with two forms of IBD, CD, or UC. As well as evaluating overall transcriptome differences among sub-types, we evaluate the disease specificity of whole blood eQTL effects in order to infer what fraction of risk can be attributed to differences in genetic regulation of gene expression.

2.2 Methods

2.2.1 Cohorts

In total, there were 190 patients and 12 controls. Protocols including signed consent of all participants and/or assent of parents in the case of minors were approved by the IRBs of Emory University and Georgia Institute of Technology. All patient cohorts were comprised of individuals of European (n = 141) or African (n = 49) ancestry from the USA. The cohorts are further divided into IBD and JIA subgroups. Within the IBD subgroup, 60 individuals were CD patients while 15 were UC patients. The average age of disease onset for CD and UC patients was approximately 14 years, with ages of onset ranging from less than 1 to 26 years. The JIA subgroup was comprised of 43 oligoarticular, 46 polyarticular, and 26 systemic JIA patients. The average age of disease onset for JIA patients was 8 years, with onset ages ranging from 0.7 to 17 years.

2.2.2 RNA-Seq processing and differential gene expression analysis

RNA was isolated from whole blood, and RNA-Seq was used to determine profiles of gene expression. The paired-end 100 bp reads were mapped to human genome hg19 using TopHat2 (120) with default parameters, with 90.4% success rate. The aligned reads were converted into number of reads per gene using SAMtools and HTSeq with the default union mode (121, 122). The raw counts were then processed by trimmed mean of M-values normalization via the edgeR R package into normalized counts (123). To further normalize and remove batch effects from gene expression data, surrogate variable analysis (SVA) combined with supervised normalization was used (124). First, FPKM was calculated and

all genes with greater than 10 individuals with greater than six read counts and FPKM > 0.1 were extracted. Expression of the sex-specific genes RPS4Y1, EIF1AY, DDX3Y, KDM5D, and XIST was used to verify the gender of each individual. The SVA R package (124) was used to identify 15 latent confounding factors, and these were statistically removed without compromising known disease variables using the supervised normalization procedure in the SNM R package (125). Pairwise comparisons between control, CD, UC, oligoarticular JIA, polyarticular JIA, and systemic JIA were performed to quantify the extent of differential expression. Using edgeR's generalized linear model likelihood ratio test function, the log fold change and Benjamini-Hochberg adjusted p value were obtained for all genes within each contrast (123).

Gene ontology analysis was performed using the GOseq R package, which incorporates RNA-Seq read length biases into its testing (126). Genes with an edgeR-calculated FDR of < 0.01 were considered to be differentially expressed and input into the GOseq software. Genes were distinguished by positive and negative log fold change to classify upregulation in specific sub-types. Only pathways within the biological processes and molecular function gene ontology branches were called.

Analysis of established immune-related gene sets was performed using BIT (Blood Informative Transcript) and BTM (Blood Transcript Module) gene expression (127, 128). The BITs are highly co-regulated genes which define seven axes of blood immune activity that are highly conserved across whole blood gene expression datasets. Standard PCA analysis including multiple PC captures most of the variance also described by the BIT, but it does so in a study-specific manner in which the actual PC have little biological

meaning. By contrast, the BIT axes, as originally characterized by Preininger et al. (127), capture components of variation that are consistently observed across all peripheral blood gene expression studies, for the most part independent of platform. We simply take PC1 for the representative genes for each axis and note that this typically explains upwards of 70% of variance of those transcripts, so it is highly representative of overall gene expression in the axis. Whereas in previous work (127) we labelled nine axes BIT axis 1 through 9, subsequent analyses and comparison with BTMs has led to affirmation of the immunological functions captured by six of the axes, which we here rename reflecting these functions as axis T (T cell-related, formerly 1), axis B (B cell-related, formerly 3), axis N (neutrophil-related, formerly 5), axis R (reticulocyte-related, formerly 2), axis I (interferon-responsive, formerly 7), and axis G (general cellular biosynthesis, formerly 4). axis 6 remains of uncertain function, while axes 8 and 9 are dropped since they are derivative and less consistent. Finally, a newly identified axis C captures numerous cell cycle-related aspects of gene activity. Each of these axes clusters with a subset of the 247 BTMs identified by Li et al. in their machine-learning meta-analysis of 30,000 peripheral blood gene expression samples from over 500 studies (128), and these relationships were visualized by hierarchical cluster analysis performed using Ward's method in SAS/JMP Genomics (129).

2.2.3 SNP data processing and eQTL analysis

The Affymetrix Axiom BioBank and Illumina Immunochip arrays were used to perform genotyping, at Akesogen Inc. (Norcross, GA). Quality control was performed using PLINK, with parameters set to remove non-biallelic variants, SNPs not in Hardy-

Weinberg equilibrium at $P < 10^{-3}$, minor allele frequency < 1%, and rate of missing data across individuals > 5% (130).

The Affymetrix Axiom BioBank array, which has a coverage of 800 k SNPs, was utilized to genotype the 115 JIA samples and 27 IBD samples. The Immunochip, which includes a high density of genotypes at loci containing markers known to be associated with various autoimmune and inflammatory diseases, including CD and UC, was used to genotype the remaining IBD samples. Following QC, imputation was performed using the SHAPEIT and IMPUTE2 software in order to merge the datasets (131, 132). However, due to the nature of the Immunochip, imputation failed to generate reliable results for sites outside of the densely genotyped regions. Consequently, the eQTL analysis was initially performed independently on the JIA and IBD datasets, and then, overlapping loci significant in either study were pooled for the interaction testing. For JIA, following QC, we analyzed 109 individuals with 5,522,769 variants. For IBD, the available Affymetrix samples were merged with the remaining 27 IBD samples from the Immunochip dataset by selecting overlapping SNPs, which following QC resulted in 54 individuals with 58,788 variants in the vicinity of the 186 immune-related loci, plus the HLA complex, included on the Immunochip. In summary, 27 IBD samples were genotyped on the Affymetrix array, while 27 were typed on the Immunochip, and the remaining 21 IBD samples had expression but not genotype data.

Using the genes from the SVA and SNM adjusted expression data and the separate compiled variants from JIA and IBD, a list of genes and SNPs within 250 kb upstream and downstream of the stop and start coordinates of the gene was generated. eQTL mapping

was performed using the linear mixed modelling method in GEMMA (133), which generated a final file of 16,913,152 SNP-gene pairs for JIA samples and 338,005 SNP-gene pairs for IBD samples. Since there are on average close to five candidate genes per SNP, between the two diseases, 263,575 SNP-gene pairs were shared that were analyzed jointly. A common p value threshold of $p < 0.0001$ corresponding to an empirical FDR < 5% was chosen, yielding 814 SNP-gene univariate associations. Conditional analysis was underpowered to detect secondary signals consistently, so we simply retained the peak eSNP associations defining 142 eGenes. Since low minor allele frequencies can drive spurious eQTL signatures if the minor homozygotes have outlier gene expression, we checked for an overall relationship between MAF and eQTL significance. None was observed, implying that rare variants are not driving the results in general, but we also examined each of the loci with significant interaction effects manually, identifying a small number of false positives. A notable example is *IL10*, which had an anomalously high disease-by-interaction ($p \sim 10^{-7}$) driven by a large effect size in IBD (beta = 2.7) that turns out to be due to a single outlier, removal of which abrogates any eQTL effect at the locus (also consistent with the blood eQTL browser report (32)).

The eQTL×disease interaction effect which evaluates whether the genotype contribution is the same in JIA and IBD was modeled by combining the imputed rsID genotypes for the lead SNP in either disease into a joint linear model with gene expression as a function of genotype, disease, and genotype-by-disease interaction, assuming the residuals are normally distributed with a mean of zero. A caveat to this analysis is that the lead SNP (i.e., the one with the smallest p value) is not necessarily the causal variant, and

secondary SNPs in one or other condition may skew the single-site evaluations. Post hoc analyses revealed that secondary eQTLs are evident at three loci reported (*PAM*, *SLC22A5*, and *GBAP1*).

2.2.4 Adjustments for medication and disease duration

Because the JIA patients in our study were not recruited from a single cohort, therapeutic interventions and duration of disease vary between individuals. Environmental factors include exposure to medications and impact gene expression profiles (134). In addition, it has previously been shown that gene expression networks are altered over the first 6 months of therapy for JIA patients (135). To characterize the effects of these covariates, our JIA patients were classified by three non-exclusive categories of medication: known treatment with disease-modifying anti-rheumatic drugs (DMARDs), biologics, and steroids at the time of sample collection, as well as three categories of disease duration prior to sampling: less than 180 days, 180–360 days, and greater than 360 days. Nearly all IBD patients were sampled at diagnosis, so this stratification was only necessary for JIA patients. Medication and time variables were then modeled and removed using SNM, resulting in an adjusted gene expression dataset (125). The previously described BIT axis analysis was performed again using this adjusted dataset and compared with results from the unadjusted dataset (Appendix B: Supplementary Figure 1A). Appendix B: Supplementary Figure 1B shows the correlation between unadjusted gene expression and category of disease duration. In addition, the JIA eQTL study was rerun using the adjusted expression dataset. The correlation of betas from the unadjusted and adjusted analyses is depicted in Appendix B: Supplementary Figure 2.

Furthermore, we were able to replicate the major trends in gene expression observed in our dataset in a published Affymetrix microarray study of samples from the various subsets of JIA (135). They studied PBMC gene expression for 29 controls, 30 oligoarticular, 49 polyarticular, and 18 systemic JIA patients all obtained prior to initiation of therapy (135). As shown in Appendix B: Supplementary Figure 3, axes R, B, N, I, and C give very similar results whereas the T cell signature which is mildly reduced in more severe JIA in our data does not differentiate their sample types. Additionally, axis G reverses the sign of effect, as it does upon adjustment for medication usage, reinforcing the conclusion that general cellular metabolic processes are affected by medication. By contrast, Hu et al. (136) report effects of anti-TNF biologic therapy specifically on certain neutrophil-related pathways, a result not recapitulated in our data, likely due to differences in experimental design.

2.2.5 Colocalization and transcriptional risk score (TRS) analysis

Colocalization analysis was performed using JIA and IBD eQTL data and prior IBD, rheumatoid arthritis, and JIA GWAS study data. The coloc R package uses a Bayesian model to determine posterior probabilities for five hypotheses on whether a shared causal variant is present for two traits (137). The analysis considered all SNPs associated with IBD (n = 232), RA (n = 101), or JIA (n = 28) as discovered by GWAS, where n = 198, 57, 21 and n = 198, 83, 20 were present in SNP-gene eQTL datasets for IBD and JIA, respectively. Cross-comparisons between both of the eQTL datasets and each of the GWAS studies' reported loci was performed, following which select SNP-gene pairs with high probabilities of hypothesis 3 (same locus but different eQTL and GWAS peaks) and 4

(same causal variant driving the signal at the eQTL and GWAS peaks) were plotted using LocusZoom (138) to visualize the region surrounding the variants.

Two independent transcriptional risk scores (TRS) were generated using GWAS results for IBD (2) and RA (139) as a proxy for JIA (since the JIA pool of variants is currently too small). As previously described, TRS sums the z-scores of gene expression polarized by the direction of effect of the eQTL relative to the GWAS risk allele (8). Thus, if the risk genotype is associated with decreased expression, we invert the z-score in the summation such that positive TRS represents elevated risk. We only used genotypes that are validated as both eQTL and GWAS by H4 in the coloc analysis, taking the eQTL list from the blood eQTL browser since it has much higher power than the small disease samples. Thirty-nine and 23 genes were included in the IBD and RA TRS, respectively, as listed in Appendix A: Supplementary Table 1. ANOVA was performed between groups to establish whether the TRS can be used to predict disease from blood gene expression.

2.3 Results

2.3.1 Heterogeneity of gene expression within and among disease sub-types

In order to contrast the nature of differential gene expression between three sub-types of JIA and two sub-types of IBD as well as relative to healthy controls, we conducted whole blood gene expression profiling on a combined sample of 202 children with disease onset between the ages of 0.7 and 17. The sample included 43 cases of oligoarticular JIA, 46 of

polyarticular JIA, 26 of systemic JIA, 60 of Crohn's disease, and 15 of ulcerative colitis. RNA-Seq analysis was performed with a median of 19.6 million paired-end 100 bp reads per sample. After normalization and quality control as described in the "Methods" section, a total of 11,614 genes remained for analysis.

Previous microarray-based gene expression profiling of JIA has established significant mean differences among disease sub-types, as well as heterogeneity within sub-types (89-92). A heat map of two-way hierarchical clustering of all genes in all individuals reveals six major clusters of individuals (rows in Fig. 7a) who share co-regulation of at least nine sets of genes (columns). For example, the top cluster labeled in dark blue consists of individuals with generally high innate immunity gene expression and low lymphocyte gene expression, whereas the bottom two clusters labeled in pale blue and green have the opposite profile, though with differences in T cell-related expression. Individuals in each of the six health and disease categories are dispersed throughout the matrix but with highly significant tendencies for enrichment of specific expression clusters in each sub-type, as shown in Fig. 7b. Eighty percent of the healthy controls are in the pale green cluster, which accounts for just one quarter of the oligo-JIA sub-type and less than 15% of each of the others. The two IBD sub-types are more likely to be in the dark blue cluster, as are sJIA cases, consistent with these being more inflammatory conditions, but in each case, the majority of individuals from each disease sub-type are dispersed throughout the other clusters. JIA in general has high membership in the red cluster, while there is an apparent gradient with oligo-JIA more control-like and sJIA more IBD-like. As with other autoimmune diseases, although there are certainly disease-related trends, the overall blood

gene expression pattern is dominated by heterogeneity without ambiguous separation by disease type. Figure 7c shows that 9.5% of the gene expression captured by the first five principal components is among disease categories and another 7.3% among the sub-types within JIA and IBD, with a small component also attributable to age-of-onset less than 6.

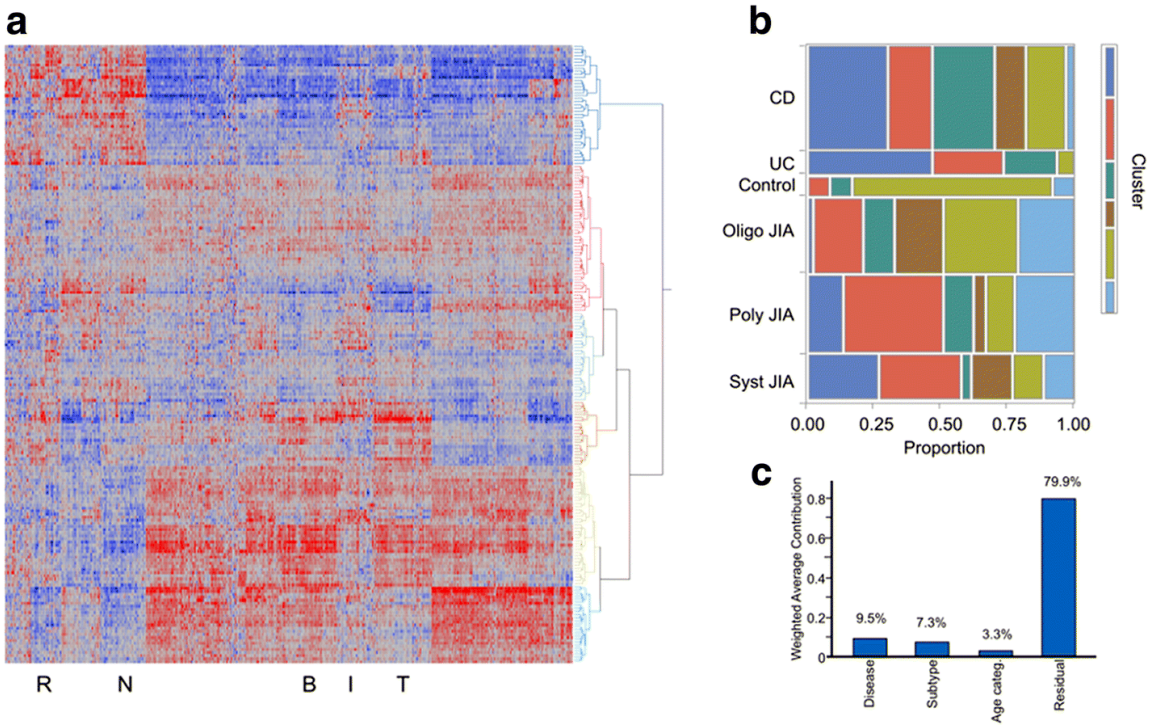


Figure 7 – Heterogeneity of gene expression within and among disease sub-types. (a) Two-way hierarchical clustering using Ward's method of standardized normal (z-scores) of transcript abundance of 11,614 genes (columns) in 202 individuals (rows). Six clusters identified to the right group individuals with similar profiles with respect to at least nine clusters of co-expressed genes. Letter beneath the heat map highlight BIT corresponding to genes enriched in reticulocytes (R), neutrophils (N), B cells (B), T cells (T), or for the interferon response (I). (b) Proportion of individuals of each disease sub-type represented in each of the six clusters of individual. For example, 45% of the UC samples are in the dark blue cluster, 30% in the red, 20% in the green, and 5% in the pale green, with none in the brown or light blue. (c) Principal variance component analysis shows the weighted average contribution of disease, sub-type within disease, or age-of-onset before 6 to the first five PC (67%) of the total gene expression variance, with the remainder residual variance unexplained, including individual differences.

2.3.2 Functional characterization of the gradient of differential expression

Contrasts of significant differential expression performed between healthy controls and sub-types of JIA as well as combined IBD and sub-types of JIA confirm the gradient of differential expression between disease groups of different severities. Appendix A: Supplementary Table 2 lists the significantly differentially expressed genes at the 5% Benjamini-Hochberg false discovery rate, for each comparison of two disease groups from the six under consideration. In the comparison between healthy controls and oligoarticular JIA, 82 genes were significantly upregulated in healthy controls, and 7 were upregulated in oligoarticular JIA. These numbers are lower than the 136 and 36 differentially expressed genes found in the contrasts between healthy controls and polyarticular JIA, and the 216 and 547 upregulated genes found between healthy controls and sJIA. A similar graded pattern of differentiation was found in comparisons of IBD and JIA. The fewest differentially expressed genes were found in the contrast between IBD and sJIA, with 73 upregulated genes in IBD and 170 upregulated genes in systemic JIA. Between IBD and polyarticular JIA, 934 upregulated IBD genes and 767 upregulated polyarticular genes were discovered, while the biggest differentiation was observed between IBD and oligoarticular JIA, where 2038 upregulated IBD genes and 1751 upregulated oligoarticular genes were discovered. These patterns of differential expression also confirm that of the three JIA sub-types, systemic JIA is the most similar to IBD.

The biological meaning of these differentially expressed genes was investigated through gene ontology and modular analysis. Contrasts between healthy controls and JIA subtypes implied a variety of classes of differential pathway regulation. Overall, all

subtypes of JIA showed downregulation of transmembrane signaling and G-protein-coupled receptor activity. However, oligoarticular JIA showed primarily upregulation of protein and phospholipid metabolic processes while polyarticular JIA showed upregulation in secretion, exocytosis, and granulocyte activation, as well as neutrophil activation. Systemic JIA showed an even more strongly significant upregulation of immune pathways, notably general immune response and myeloid activation. In contrast, for the comparisons between IBD and JIA subtypes, all JIA subtypes showed upregulation of nucleic acid processes compared with IBD. Both oligoarticular and polyarticular JIA showed strongly significant downregulation of myeloid, neutrophil, and leukocyte activity compared with IBD, whereas sJIA showed downregulation of general metabolic processes albeit at a much lower significance level.

2.3.3 Clustering by BTMs and BITs further reveals enriched immune pathways

Decades of blood gene expression analysis have highlighted the existence of modules of co-expressed genes that reflect a combination of joint regulation within cell types and variable abundance of the major leukocyte classes (140). Seven highly conserved axes of blood variation (127) are composed of genes broadly capturing immune activity related to T and B cells, reticulocytes and neutrophils, interferon response, general biosynthesis, and the cell cycle. Figure 8 shows clear trends of expression along these axes correlating with disease sub-type, each panel indicating the level of activation in each immune component in, from left to right, healthy control, oligoarticular JIA, polyarticular JIA, systemic JIA, Crohn's disease, and ulcerative colitis. Axis T, representing T cell expression, and axis B, representing B cell expression, show a trend of decreasing PC1

values correlating with severity of disease, suggesting downregulation of adaptive immunity in systemic JIA, CD, and UC. In contrast, axis R, representing reticulocytes, and axis N, representing neutrophils, show trends of increasing PC1 values with disease severity that indicates upregulation of the innate immune system in systemic JIA, CD, and UC. Axis I represents interferon-responsive gene expression and has a more parabolic trend, being elevated in polyarticular and systemic JIA and Crohn's disease, but not ulcerative colitis, reflecting the interferon response's dual roles in both adaptive and innate immunity. Axes G and C represent general and cell cycle expression, and show trends of higher PC1 values in inflammatory bowel disease and systemic JIA. Despite sample sizes of around 30 patients in each group, ANOVA indicates that the differences are significant in each case.

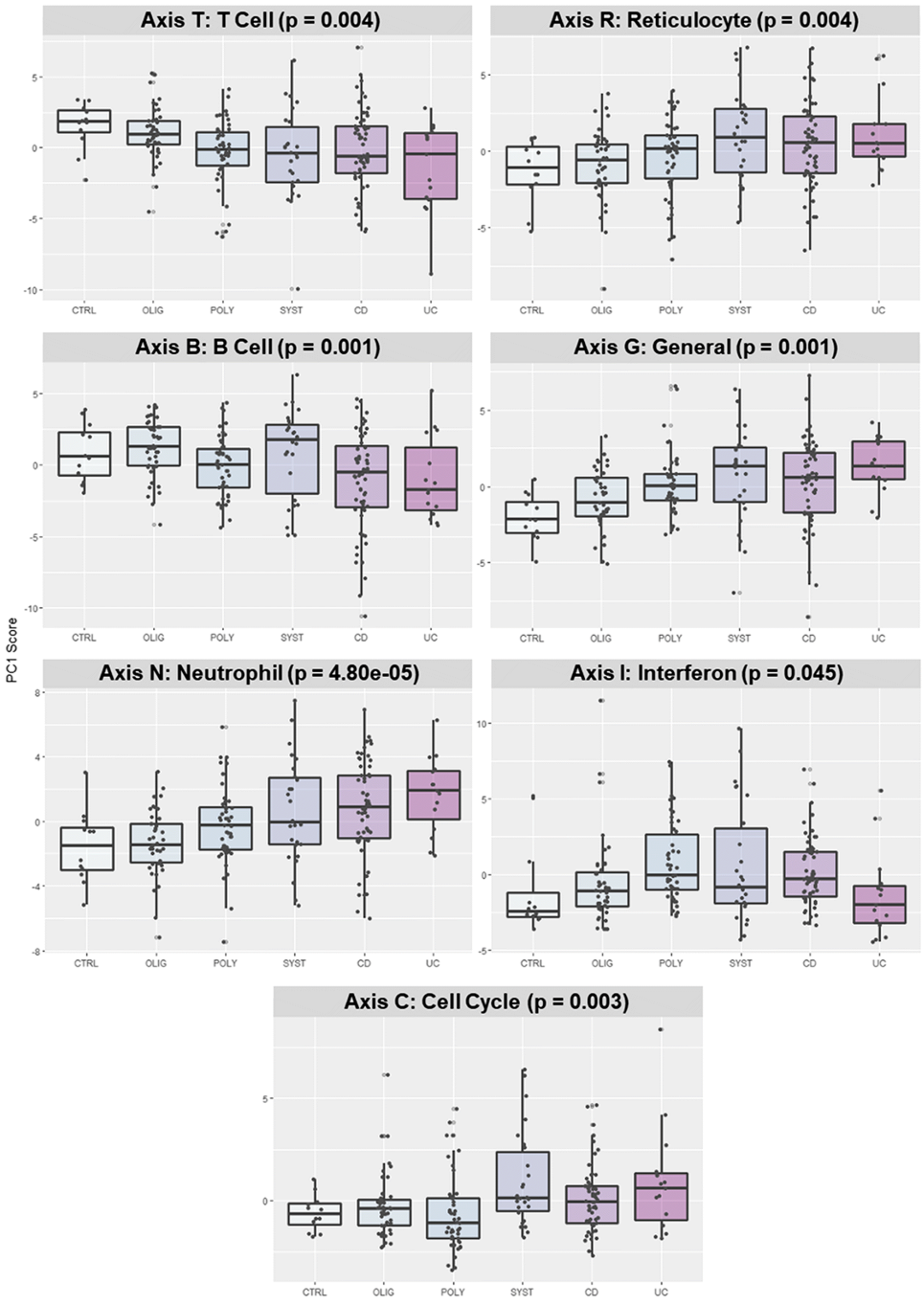


Figure 8 – Axes of variation across disease sub-types. Axes of variation defined by the first PC of the Blood Informative Transcripts (BIT) highlight variation in types of immune activity across disease sub-types. Each individual data point represents PC1 score for 10 BIT for the indicated axis, with box and whisker plots showing the median and interquartile range as well as 95% confidence intervals for the sub-types. Indicated p values are from one-way ANOVA contrasting the six sub-types of sample.

These disease-specific trends are confirmed by hierarchical clustering of 247 Blood Transcript Modules (BTMs) (128) in Fig. 9, tabulated in Appendix A: Supplementary Table 3, further supporting the gradient of disrupted gene expression based on disease severity. Healthy controls and oligoarticular JIA show largely similar expression, except for apparent elevation of NK cell gene expression in controls. IBD most resembles sJIA, although with some key differences. Myeloid gene expression tends to be elevated in IBD and lymphoid gene expression suppressed, with JIA intermediate. In addition, ulcerative colitis appears to have a specific deficit in NK cell-biased gene expression, sJIA has a unique signature including inositol metabolism, and JIA in general shows reduced mitochondrial gene activity.

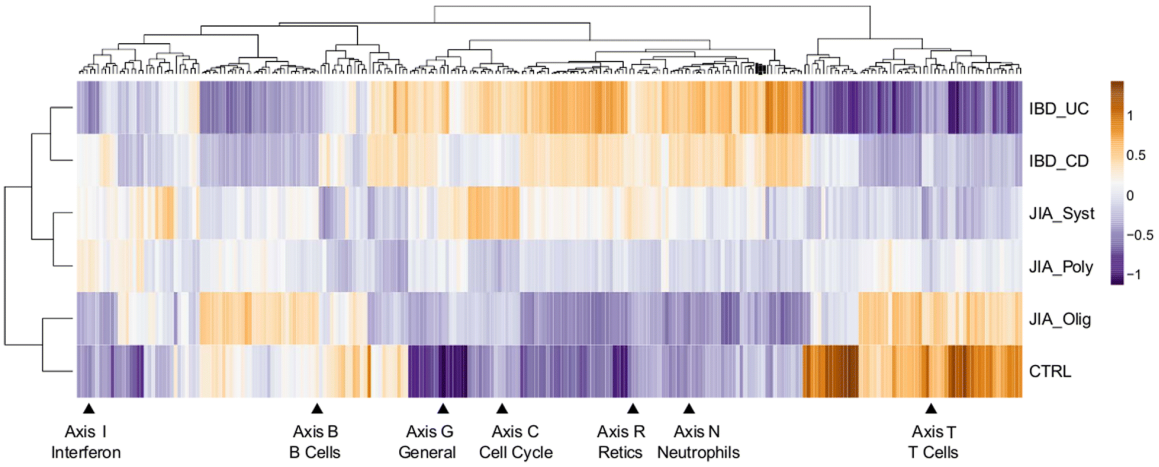


Figure 9 – Blood Transcript Modules. Hierarchical clustering of blood transcription modules across disease sub-types. The heat map shows the mean PC1 scores for 247 BTM identified in (127), as well seven BIT axes. Note how the BTM form ~ 10 clusters, seven of which co-cluster with one orthogonally determined axis. See Appendix A: Supplementary Table 3 for a complete listing of BTM scores in each disease sub-type.

2.3.4 Transcriptional risk scores differentiate healthy controls, JIA, and IBD

We recently proposed the notion of a transcriptional risk score (TRS), which is analogous to a cumulative burden of genotypic risk, but evaluates cumulative burden of risk due to elevated or suppressed gene expression relevant to disease (8, 83). By just focusing on genes with shared eQTL and GWAS associations, the analysis is restricted to genes most likely to have a causal role in pathology, whether because the risk allele directly promotes disease or fails to provide sufficient protection. A TRS based on eQTL detected in blood but with gene expression measured in ileum was highly predictive of Crohn's disease progression, whereas a corresponding genetic risk score was not. Figure 10 shows similarly that the 39-gene IBD TRS measured in peripheral blood provides significant discrimination of cases and controls (difference in standard deviation units of TRS; Δs.d. = 1.10, p = 0.0003); notably, sJIA is elevated to the same degree as both CD and UC. By contrast, oligoarticular JIA and polyarticular JIA have intermediate TRS that are nevertheless significantly greater than healthy controls (Δs.d. = 1.04, p = 0.0031). For comparison, a TRS based on genes that are likely to be causal in driving the signal at 23 genome-wide significant associations for RA does not discriminate between healthy controls and IBD as a group (Δs.d. = 0.11, p = 0.63) but does trend toward discrimination of JIA as a category (Δs.d. = 0.42, p = 0.09). This RA TRS is mostly enhanced in sJIA (Δs.d. = 0.86, p = 0.008 relative to healthy controls), suggesting that it is capturing the effects of inflammatory gene contributions to this most severe form of JIA.

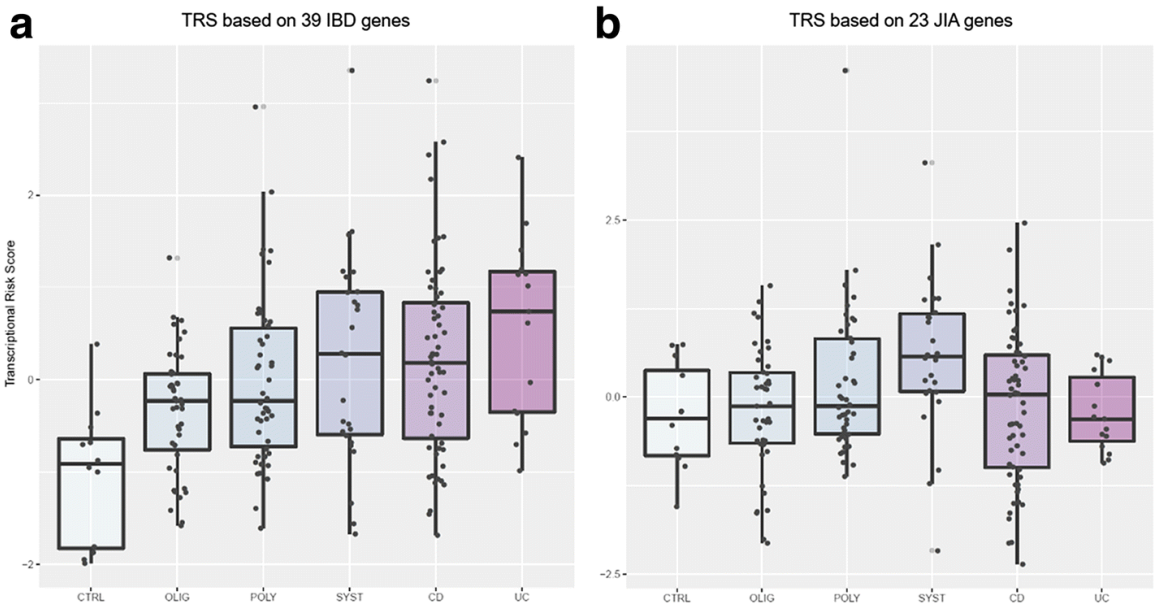


Figure 10 – Transcriptional risk scores associate with disease status. (a) IBD-TRS scores within disease sub-types for 39 genes associated with IBD in (2). Gene expression values for each selected gene were transformed into z-scores, polarized relative to risk according to whether the eQTL activity of the risk allele discovered by GWAS increases or decreases transcript abundance, and summed to generate the TRS as in (8). (b) New RA-TRS based on 23 genes associated with RA by GWAS (139).

2.3.5 Evaluation of disease specificity of eQTL

We next addressed the degree of sharing of the local genetic control of gene expression in the two classes of disease (namely JIA and IBD) by performing comparative eQTL analysis. Whole genome genotypes were ascertained on the Immunochip (CD and UC samples) or the Affymetrix Axiom Biobank array (see the "Methods" section). As far as possible, SNPs were imputed onto the 1000 Genomes reference, allowing cross-comparison of the disease subsets, noting that this was not possible for loci not included

on the Immunochip. Since genotypes were generated on different platforms, the eQTL assessment was first performed independently for the two broad disease classes, after which significant effects were evaluated jointly. Here we only consider genes located within the vicinity of the Immunochip loci.

For JIA, 107 independent eSNPs were identified within 500 kb of a transcript at an FDR of 5% (approximate $p < 10{-}4$), and for IBD, which had a smaller sample size, 52 independent eSNPs were identified. These are listed in Appendix A: Supplementary Table 4. Twelve of the loci overlap between the two diseases, but failure to detect an eQTL in one condition does not necessarily imply absence of the effect, since the small sample size results in relatively low power. Overall, the correlation in effect sizes is high, ~ 0.7 ($p = 5 \times 10^{-20}$ in JIA; $p = 2 \times 10^{-8}$ in IBD), which is remarkable given the small sample sizes, and strongly implies that most eQTL effects in whole blood are consistent across the diseases. Nevertheless, the plots in Fig. 11 depicting the estimated eQTL effect sizes in IBD relative to JIA provide some support for disease-biased effects in so far as the eQTL discovered in JIA (red points, panel a) tend to have larger effects on JIA (beta values) than those observed in IBD and hence lie between the diagonal and the x-axis. Conversely, the eQTL discovered in IBD (blue points, panel b) tend to have larger effects on IBD than those observed in JIA and hence lie between the diagonal and the y-axis. This result is biased by winner's curse, the tendency to over-estimate effect sizes upon discovery, so we also evaluated all associations jointly in order to also identify interaction effects. At an FDR of 10%, 34 of the 147 independent eQTL, highlighted in panel c, show nominally significant interaction effects ($p < 0.02$), implying different effect sizes in the two broad

classes of disease. Example box plots of genotypic effects on transcript abundance across the two disease classes are provided in Appendix B: Supplementary Figure 4. These genotype-by-disease interaction effects remain significant after accounting for ancestry (see Appendix B: Supplementary Figure 5).

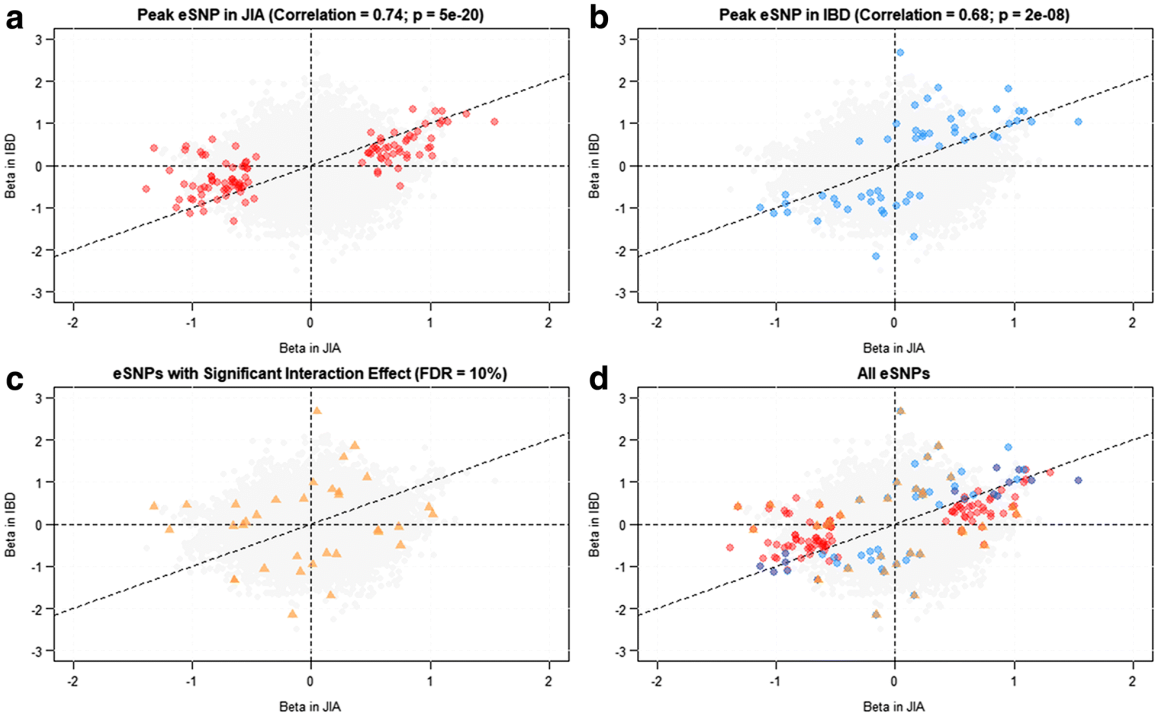


Figure 11 – Comparison of peripheral blood eQTL effects between JIA and IBD. Effect sizes of peak eSNPs by disease. (a) Correlation of beta effect sizes between IBD and JIA for the 107 peak independent eSNPs discovered in the JIA sample. (b) Correlation of beta effect sizes between IBD and JIA for the 52 top eSNPs identified in JIA. (c) Thirty-four eSNPs with a significant interaction effect between disease and genotype when evaluated jointly. (d) Overlay of all eSNPs.

As expected, many of the detected eQTLs affect expression of genes in the vicinity of established GWAS hits for autoimmune disease. Table 1 lists 25 lead eSNPs that regulate expression in cis of 22 target genes that are listed on ImmunoBase as potential causal genes for IBD or arthritis (JIA or RA). Half of these associations are with IBD only, but this bias may simply reflect increased power of the IBD GWAS to date. Several of the SNPs show evidence of disease-specific or disease-biased effects. Naively, we might expect the eQTL to be seen only in the disease(s) for which the association with disease is seen, as this would be consistent with allele-specific expression driving pathology. Three cases (*ARPC2, CPTP* for IBD, and the secondary eQTL in *PAM* for JIA) fit the expected pattern, but three others have the counter-intuitive relationship where the eQTL is observed in one disease but the established GWAS association is with the opposite disease (*PRDX6* and *ADAM1A* for RA, the secondary eQTL in *GBAP1* for CD). Three more cases (*SLC22A5, CD226*, and *RNASET2*) have possibly disease-biased eQTL effects where the eQTL is absent from or much less in one disease, although the interaction effect is only significant in one of these cases. Despite the small sample, there is not an intuitive pattern to the relationship between disease-biased regulation of gene expression and association with disease.

Table 1 – 25 lead eSNPs that regulate expression in cis of 22 target potential causal genes for IBD or arthritis (JIA or RA)

Gene	rsID	IBD β	IBD *p* val	JIA β	JIA *p* val	IBD-GWAS	ATH-GWAS	Interact *p*
ARPC2	rs13429408	0.82	6.60E−05	0.18	0.22	CD, UC	–	0.01
CPTP	rs11809901	− 1.08	9.80E−05	− 0.12	0.69	CD, UC	–	0.04
PAM	rs2431321	1.04	3.80E−09	1.15	2.10E−23	–	RA	0.48
PAM	rs32677	0.21	0.3	0.94	5.30E−15	–	RA	9.60E−05
C5	rs1468673	0.39	0.02	0.74	3.10E−07	–	RA	0.34
PRDX6	rs4279882	1.84	3.80E−05	0.36	0.05	–	RA	0.001
ADAM1A	rs11066027	1.22	2.40E−05	0.61	5.30E−03	–	JIA, RA	0.09
RNASET2	rs385863	− 0.68	1.30E−04	− 1.05	1.40E−14	CD, UC	RA	0.3
GSDMB	rs11078926	− 0.51	5.90E−03	− 0.56	9.90E−07	CD, UC	RA	0.87
SLC22A5	rs11739135	0.09	0.6	− 0.8	9.80E−10	CD, UC	JIA	4.00E−05
SLC22A5	rs11950562	− 0.53	8.00E−04	− 0.86	6.10E−14	CD, UC	JIA	0.07
ORMDL3	rs1565923	1.11	8.80E−07	0.47	6.20E−04	CD, UC	RA	0.01
ICAM4	rs3093029	1.22	4.80E−04	1.3	2.90E−08	CD, UC	JIA	0.69
RMI2	rs11644184	− 0.58	7.60E−04	− 0.7	3.00E−07	CD, UC	JIA	0.54
PLTP	rs7275164	− 0.56	2.10E−04	− 0.71	7.00E−07	CD, UC	RA	0.58
CD226	rs12969613	0.63	2.20E−07	0.18	0.15	CD, UC	RA	0.11
NOD2	rs1981760	1.28	2.70E−08	1.05	2.30E−16	CD	–	0.23
GBAP1	rs914615	0.6	3.20E−04	0.8	7.80E−10	CD	–	0.62
GBAP1	rs3814319	0.16	0.33	0.7	1.20E−06	CD	–	0.05
KSR1	rs2945378	− 0.48	6.20E−03	− 0.6	4.40E−07	CD	–	0.52
SULT1A1	rs7191548	− 0.49	6.50E−03	− 0.61	5.30E−07	CD, UC	–	0.93
PNKD	rs13430006	0.34	0.14	0.57	6.80E−07	CD, UC	–	0.41
NLRP2	rs12975582	0.56	0.01	0.8	1.20E−06	CD, UC	–	0.43
SLC11A1	rs78846874	− 0.35	0.36	− 0.83	3.90E−06	CD, UC	–	0.22
LGALS9	rs1984547	− 0.88	2.40E−05	− 0.55	4.10E−05	CD, UC	–	0.16

One reason for divergent effect sizes may be that different causal variants in variable degrees of linkage disequilibrium could be responsible for the differential expression in the two disease sub-types. To investigate this, we performed colocalization analysis using coloc (137) to visualize the locus-wide SNP effects across all loci reported in IBD, RA, and JIA GWAS and present in our SNP-gene datasets for IBD or JIA and compared these with the distribution of GWAS summary statistics. Coloc assigns a posterior probability that the same SNP is responsible for both an eQTL effect and the disease association (H4) or that different SNPs are responsible for the two effects (H3). Since the power of this mode of analysis is limited when sample sizes are small, we identified cases from either disease with relatively strong H3 or H4 posterior probabilities and plotted representative examples in Fig. 12. The full results are summarized in Appendix A: Supplementary Table 5.

Figure 12a shows results for association of rs12946510 with IBD from GWAS (bottom panel) and the eQTL profiles for the JIA (top panel) and IBD (middle panel) gene expression. Although coloc calls both cases as H4, the correspondence of SNP profiles in high LD with the lead SNP is more notable in JIA. The light blue SNPs suggest a second, independent, eQTL which does not produce a GWAS signal. Hence, the gene expression difference may be mediated by two different SNPs, possibly with different effect sizes in the two diseases, only one of which appears to contribute strongly to disease risk. Figure 12b shows a clear H3 case in JIA where the eQTL effect on expression of *PAM* appears to be mediated by a cluster of variants to the left of the lead GWAS cluster. Figure 12c shows a classical H4 where the fine mapping supports a single causal locus for both the gene

expression and disease, although the precise identity of the causal variant is impossible to ascertain from the statistical data alone owing to the extensive block of variants in high LD.

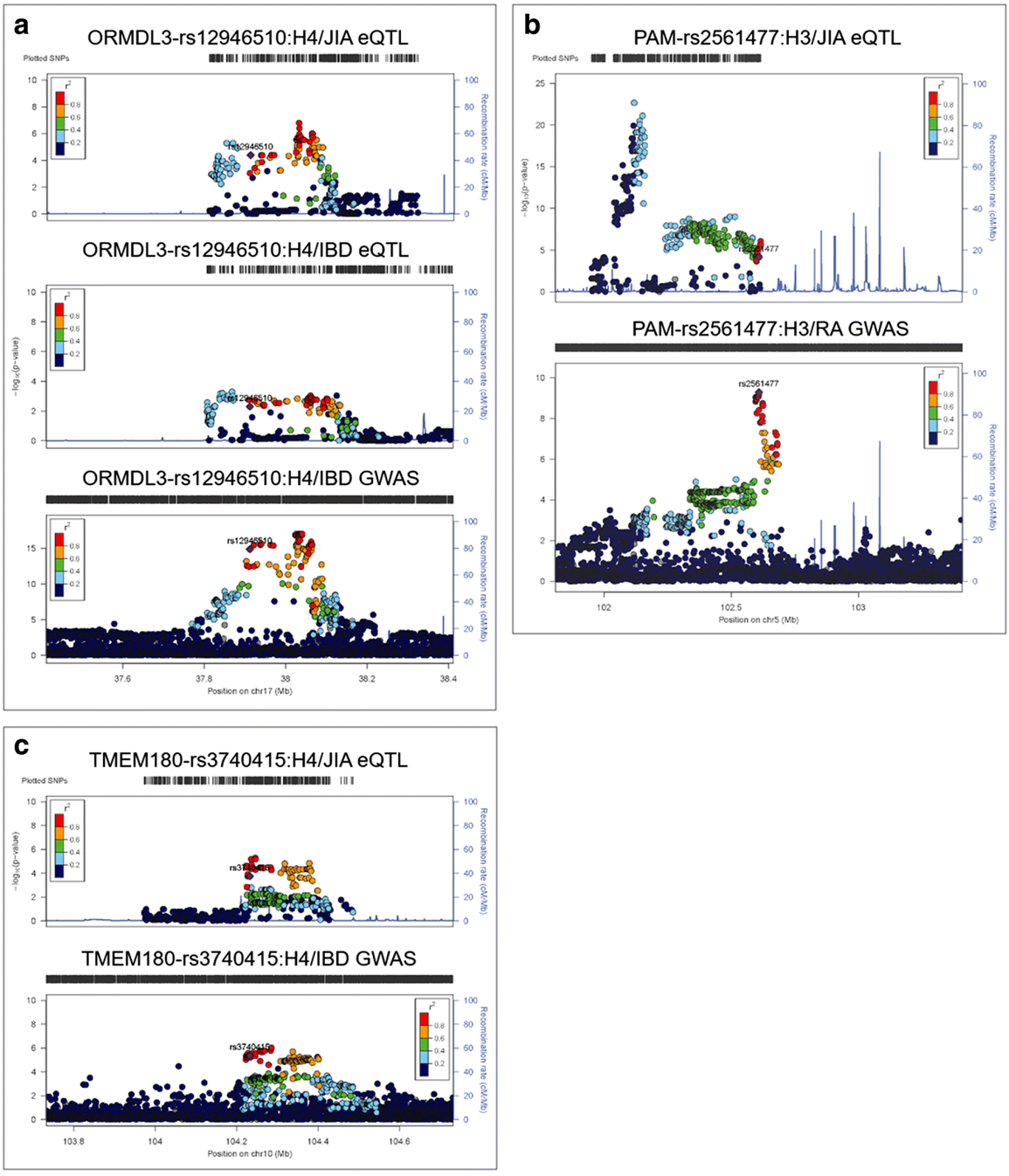


Figure 12 – Colocalization of eQTL and GWAS signatures. LocusZoom plots show the univariate SNP-wise association statistics for each genotyped SNP either with the abundance of the indicated transcript (eQTL effects) or from the GWAS for IBD or RA. Color coding indicates the r2 measure of linkage disequilibrium of each SNP with the relevant peak GWAS SNP. (a) rs12946510 is most likely a shared causal variant for *ORMDL3* gene expression in both IBD and JIA, as well as in the IBD GWAS. However, a likely secondary signal in the light blue region is not associated with IBD. (b) rs2561477 is the peak causal variant in RA but clearly does not colocalize with the peak eQTL for JIA. (c) rs3740415 is most likely a shared causal variant for expression of *TMEM180* and in the IBD GWAS despite an extensive LD block at the locus (though it does not meet the strict GWAS threshold).

2.4 Discussion

2.4.1 Disease-specific associations with autoimmune disease

There are multiple technical reasons why GWAS may fail to detect associations that are shared across multiple autoimmune diseases. These include differences in sample size and clinical heterogeneity, and with respect to eQTL analysis, differences in expression profiling platform, statistical methodology, and effects of pharmacological interventions could all obscure associations. However, it is also clear that the genetic correlation across diseases is significantly less than one, establishing the expectation that some effects must be disease-specific (141). The most appropriate framework for detecting such effects is evaluation of the significance of genotype-by-disease interaction terms, which motivated the current study.

The core result of the comparative eQTL component of this study is that the majority of genetic influences on transcript abundance measured in whole blood are consistent across IBD and JIA. A major caveat to this conclusion is that immune cell sub-

type specific effects will often go undetected in both whole blood and PBMC studies (97, 101). It is though important to note that while neutrophils, lymphocytes, macrophages, and monocytes certainly do have unique and disease-relevant eQTL, comparative studies also confirm that over three quarters of eQTL are shared by the majority of immune cells (142, 143).

Just as importantly, equivalence of genetic influences on gene expression does not necessarily mean equivalence of genetic influences on disease susceptibility. Among the shared eQTL, some genes are still likely to be specific to CD, UC, JIA, or other conditions by virtue of other influences. These may include disease-specific contributions of the critical cell type, environmental differences (for example, microbial infection of the gut may elevate or suppress expression of the gene to a degree that renders the eQTL meaningful or irrelevant), or interactions with the genetic background (for example, elevated expression of a gene may only matter in the context of other genetic risk factors). Although there is little evidence that two-locus genotype-by-genotype interactions contribute meaningfully to heritability (144), renewed interest in influences of overall genetic risk on the impact of specific genotypes makes sense given the context of gene expression heterogeneity (145).

Our analyses do provide evidence that as many as 20% of eQTL effects in peripheral blood may at least show disease-specific biases. Such differences in effect sizes are likely to trace to differences in the expression of transcription factors and epigenetic modifications between diseases and/or to differences in the relative abundance of contributing cell types. Methods exist for deconvoluting effects of cell-type abundance

(146), but they are low resolution and in our opinion unreliable when applied to sample sizes of the order of 100; next-generation studies incorporating single-cell RNA-Seq will be much more informative.

The relationship between disease-specific eQTL and GWAS association at the same locus is less straightforward than might be expected under the assumption that the effect of a polymorphism on disease is mediated through its effect on transcription of the associated gene. It is not immediately clear why an eQTL may only be detected in one disease while the GWAS association is in another disease, yet multiple instances are found in our data. This observation adds to a growing body of data questioning whether detected eQTL effects explain causal associations. Two fine mapping studies of IBD published in 2017 (147, 148) both found less than 30% identity between mapped eQTL and GWAS causal intervals, one suggesting that there is more significant overlap with methylation QTL and both arguing that the relevant effects may be specific to particular cell types or activation conditions, including immune activity at the sight of the pathology. Additionally, we described a meaningful number of “incoherent” associations, where mean differential expression between cases and controls is in the opposite direction to that predicted by the effect of the risk allele on gene expression (8). Such results highlight the need for a combination of fine structure mapping of causal variants and detailed mechanistic studies of immune cell-type contributions if we are to fully understand how segregating polymorphisms contribute to disease susceptibility and progression.

2.4.2 Disease- and sub-type-specific gene expression

Numerous other studies have described gene expression profiles in a variety of inflammatory autoimmune diseases, but we are aware of just a single side-by-side comparison of two or more diseases on the same platform (143). Straightforward cluster analysis shows that both IBD and JIA subjects tend to differ from healthy controls, but they have overall transcriptome profiles that may belong to a half dozen types. Blood Transcript Module and BIT axis analyses, both based on comprehensive analysis of existing whole blood gene expression datasets, confirm that these types broadly reflect differences in gene activity in the major immune sub-types, partly reflecting cell abundance, but also innate states of activity of biosynthetic, cell cycle, and cytokine signaling. Immunoprofiling by flow cytometry has established that individuals have baseline profiles, or omic personalities (149), to which they return after immunological perturbation but which are also influenced by such environmental factors as child-rearing (150). Sub-type-specific blood gene expression should be seen in light of this immunological elasticity, as the heterogeneity among subjects may be more meaningful for disease risk than individual eQTL effects.

Juvenile idiopathic arthritis is the most prevalent childhood rheumatic disease, encompassing multiple physically, immunologically, and genetically different sub-types of disease. Although diagnosis and classification is based upon largely clinical criteria, the genetic complexity of JIA has been well documented (110, 111). While the oligoarticular and polyarticular sub-types demonstrate activation of adaptive immunity, systemic JIA appears to be mediated more heavily through innate immunity, and profiles of immune cell

activity between sub-types differ (91, 151, 152). These findings at the gene expression level are consistent with emerging GWAS results suggesting that systemic JIA is etiologically a quite different disease. It is particularly noteworthy that both of the transcriptional risk scores we document show that systemic JIA is divergent from the articular forms, being close to the IBD profiles for the IBD-TRS, and uniquely elevated for the RA-TRS.

In this study, we performed cross-sub-type and disease comparisons of gene expression and eQTLs to characterize the similarities and differences between the forms of JIA. Differential gene expression analysis revealed a gradient of order among the JIA sub-types and IBD, from healthy controls, to oligoarticular, polyarticular, and systemic JIA, to Crohn's disease and ulcerative colitis. Numbers of differentially expressed genes, gene ontology pathway types, and significance levels agree with this pattern of ordering. Consistent with previous research, oligoarticular and polyarticular JIA exhibits a trend of activated T cell gene expression relative to systemic JIA (100-103, 106). As a group, JIA also demonstrates increased expression of B cell-related genes. There is also an ordered increase in neutrophil gene expression from oligoarticular to systemic JIA, which concurs with systemic JIA being closely tied with innate immunity. In addition, the elevation of oligoarticular and polyarticular JIA over controls points to involvement of neutrophils in these sub-types as well, which has been previously suggested (88). Taken as a whole, these findings suggest that JIA sub-types are mediated through a complex relationship between adaptive and innate immunity, and neither disease can be fully characterized by simply one or the other.

2.4.3 Limitations

This study has three major limitations. Firstly, since the subjects were not a part of any single-cohort study, they were treated with different medications or had samples taken at later time points after diagnosis. The sample size, though larger than many published studies, is still too small to partition the effects of plausible technical covariates or of environmental mediators of gene expression such as those described by Favé et al. and Idaghdour et al. (134, 153). The results of the covariate-adjustment analyses presented in Appendix B: Supplementary Figures 1 and 2 suggest that the effects on our dataset are minimal compared with the consistent effect of disease subtype, but therapeutic effects should still be considered in interpretations of our findings. Secondly, whole blood samples were utilized to measure gene expression. Because whole blood is composed of multiple cell types, there will inherently be some mixture and dilution of gene signatures. Although it is well established that whole blood expression profiles are capable of illuminating aspects of autoimmune pathology, immune cell sub-type analyses will have higher resolution (101). Single-cell RNA-Seq has great potential both to trace general features of peripheral blood gene expression to specific cell types and to foster accurate eQTL analysis at the sub-type level. Thirdly, we describe just a cross-sectional snap shot of the transcriptome of each subject, whereas longitudinal profiling has the promise of correlating personalized transcriptional shifts to clinical response (154).

2.5 Conclusions

Gene expression and genotyping data can help to categorize sub-types of JIA and IBD beyond just clinical features. The gradient of gene expression from healthy controls to oligoarticular, polyarticular, and systemic JIA to IBD reflects a complex interplay between adaptive and innate immunity responsible for differentiation between JIA sub-types. Individuals have sub-type-specific probabilities of having one of a small number of global gene expression profiles. Since the majority of eQTL appear to have similar effect sizes across disease sub-types, disease-specific cis-eQTL effects only explain a small fraction of disease-specific genetic influences on disease. Considerably more fine mapping and functional analysis will be required before personalized therapeutic interventions for patients with distinct forms of JIA or IBD become commonplace.

CHAPTER 3. AFRICAN ANCESTRY PROPORTION INFLUENCES ILEAL GENE EXPRESSION IN INFLAMMATORY BOWEL DISEASE

3.1 Introduction

The influence of ancestry on inflammatory bowel disease (IBD) susceptibility has recently been examined via several large-scale genome-wide association studies (GWAS) (78, 79), but the effects of ancestry-specific variation in risk and modifier gene expression on prognosis in IBD remain poorly characterized. It has been estimated fairly consistently that the prevalence of IBD is approximately two-to-three times greater in individuals of European versus African American (AA) descent (53), but the literature offers conflicting reports on complications and outcomes in AA versus European ancestry patients (4, 7). AA individuals tend to be admixed, with approximately 80% African (YRI) and 20% European ancestry (CEU) (61). Although GWAS have identified hundreds of variants associated with IBD risk, most research has been conducted on cohorts of exclusively European ancestry (39). Earlier studies on IBD risk variants such as *NOD2* concluded that mutations in AA individuals result from European admixture, and thus confer similar increases in risk (74). However, the largest GWAS study to date of IBD in AA identified two novel African-specific loci associated with IBD, hinting at the existence of African-specific contributions to disease that remain to be elucidated (78).

In this chapter, I evaluated ileal transcriptomic profiles of 154 individuals of mostly AA and European ancestry with IBD, and characterized differential gene expression between populations. I then examined the effect of proportions of African and European ancestry in AA patients on gene expression, demonstrating that observed variation in gene expression between populations is heritable and not solely due to environmental differences. This study was performed in collaboration with the Kugathasan lab at Emory University, and our findings have been published in *Cellular and Molecular Gastroenterology and Hepatology* (155)

3.2 Methods

3.2.1 Cohort

In total, 129 patients and 25 controls were profiled for this study. Protocols included signed consent of all participants and/or assent of parents in the case of minors, and were approved by the IRBs of Emory University and Georgia Institute of Technology. Of the 154 total participants, 121 self-identified as African American and 33 identified as white. The cohort was evenly divided by gender, with 78 female participants and 76 male participants. Suspected IBD, chronic abdominal pain without known etiology, and unexplained weight loss were amongst the most common indications for colonoscopy to be performed in control individuals. Controls retained for this study had normal colonoscopy without inflammation, as well as normal histology verified through multiple pinch biopsies. The 129 patients in this study included 36 individuals with ulcerative

colitis, of whom 28 were African American and 8 were European ancestry, and 93 individuals with Crohn's disease, of whom 76 were African American and 17 were European ancestry. The average age of onset amongst patients with UC and CD was approximately 14 years. Amongst the characterized CD patients, 18 had L1 (ileal), 11 had L2 (colonic), 53 had L3 (ileocolonic), and 1 had L1-L4 (upper gastrointestinal disease) disease location; 54 had B1 (non-stricturing), 19 had B2 (stricturing), and 8 had B3 (penetrating) status. Amongst the characterized UC patients, 4 had E1 location (ulcerative proctitis), 6 had E2 location (distal), and 25 had E3 disease location (proximal).

3.2.2 RNA-Seq processing and gene expression analysis

RNA was isolated from biopsies of the ileum for Lexogen 3' sequencing. Single end 75bp reads were trimmed for adapters with FastQC and Trim Galore, then mapped to human genome GrCh37 with the hisat2 aligner (156, 157). The aligned reads were converted into read counts per gene using HTSeq (122). The raw read counts were normalized with the edgeR R package implementation of trimmed mean of M-values normalization (123). A combination of surrogate variable analysis (SVA) and supervised normalization (SNM) was then applied to remove batch effects and other confounding factors (124, 125). First, expression of the sex-specific genes *RPS4Y1, EIF1AY, DDX3Y, KDM5D*, and *XIST* was checked to verify reported gender, resulting in the exclusion of 13 non-matching individuals. The SVA R package was then used to identify 6 surrogate variables which were then removed via supervised normalization in the SNM R package. Pairwise differential gene expression testing between African American and white IBD patients was then performed using the voom R package, which generated log fold change

and Benjamini-Hochberg adjusted p-values for all genes (158). Hierarchical clustering of the 2,705 genes differentially expressed at FDR < 0.05 was performed with the NMF R package. Gene set enrichment analysis was performed with GSEA, using pre-ranked mode on all 14,392 genes ranked by multiplying the sign of the fold change by the inverse of the Benjamini-Hochberg adjusted p-value (159). Principal components of sets of differentially expressed genes were used to evaluate whether case-control status or therapeutic regimen explain the ancestry effects, as plotted in Appendix B: Supplementary Figure 6. With the exception of steroids, which were only given to a subset of AA patients, neither of these factors associate with ancestry.

3.2.3 Variant calling and calculation of ancestry proportion

The GATK Best Practices workflow for calling variants in RNAseq was followed to generate a VCF file of SNPs for individuals in this study (160). VCF files for 1000 Genomes individuals belonging to either the CEU population (n=85) or YRI population (n=88) were extracted (161). Both VCF files were merged, and quality control for genotyping rate was performed with PLINK, restricting the dataset to 12,819 variants (162). Ancestry proportions for African American individuals were assigned using ADMIXTURE software in supervised mode, where 1000 Genomes CEU and white individuals from this study were provided as a known European population, and 1000 Genomes YRI individuals were provided as a known African population (163). Plots of ancestry proportions were generated using the pophelper R package.

3.2.4 Calculation of heritable portion of gene expression variation

The calculation of the heritable portion of observed gene expression variation between populations in this study was based on methods first described by Price et al (164). Individuals in this study were separated into CEU+YRI and African American population groups. 33 white individuals and 33 individuals with African ancestry proportions ~ 0.9999 were grouped into the CEU and YRI categories, while all other individuals were classified as African American. Gene expression across each gene was z-score normalized in the CEU+YRI group and African American group. Expression in the CEU+YRI group can be modeled as $e_{gs}=a_g\theta_s+\nu_{gs}$, where e_{gs} represents expression of gene g in individual s, a_g represents observed gene expression differences between CEU and YRI, θ_s denotes genome-wide African ancestry of either 0 or 1, and ν_{gs} represents residual effects. Then, $e_{gs}=ca_g\theta_s+\nu_{gs}$ for the African American group, where θ_s now ranges from 0 to 1 and c is a coefficient representing the extent to which a_g is heritable. An estimate of $a_{g,CEU+YRI}$ can be obtained by regressing e_{gs} against θ_s within the CEU+YRI group, and similarly an estimate of $a_{g,AA}$ can be obtained by regressing e_{gs} against θ_s within the African American group. An estimate of c can then be obtained by regressing the two estimates of a_g. The statistical significance of the estimated c was validated by testing the values of c obtained from 1000 sets of random permutations of African ancestry among African American individuals, then ranking the correlations. The permutation test yielded a p-value of 0.05 for the c estimate based on true ancestry.

3.3 Results

We performed RNA-seq of ileal biopsies sampled from control individuals (n=25, no intestinal inflammation and normal histology) and 129 patients with ulcerative colitis (UC, n=36) and Crohn's disease (CD, n=93). Differential gene expression analysis revealed 1,360 upregulated and 1,345 downregulated genes at an FDR cutoff of 0.05 in AA patients compared with European ancestry patients (Figure 13). Hierarchical clustering based on these 2,705 genes shows separation of transcriptomic profiles by ancestry into two clusters (Figure 13b). To explore functional pathways implicated by differentially expressed genes, we performed Gene Set Enrichment Analysis (GSEA). Oxidative phosphorylation, adipogenesis, and xenobiotic metabolism were amongst the gene sets enriched for genes upregulated in AA, while TNFα signaling, inflammatory response, and interferon-γ response were enriched in downregulated genes (Figure 13c, Appendix A: Supplementary Table 6). Each of these pathways is highly relevant to the development and progression of pathology in IBD, which highlights the importance of better understanding of the genetic contributions to the disease and possible personalized treatment options.

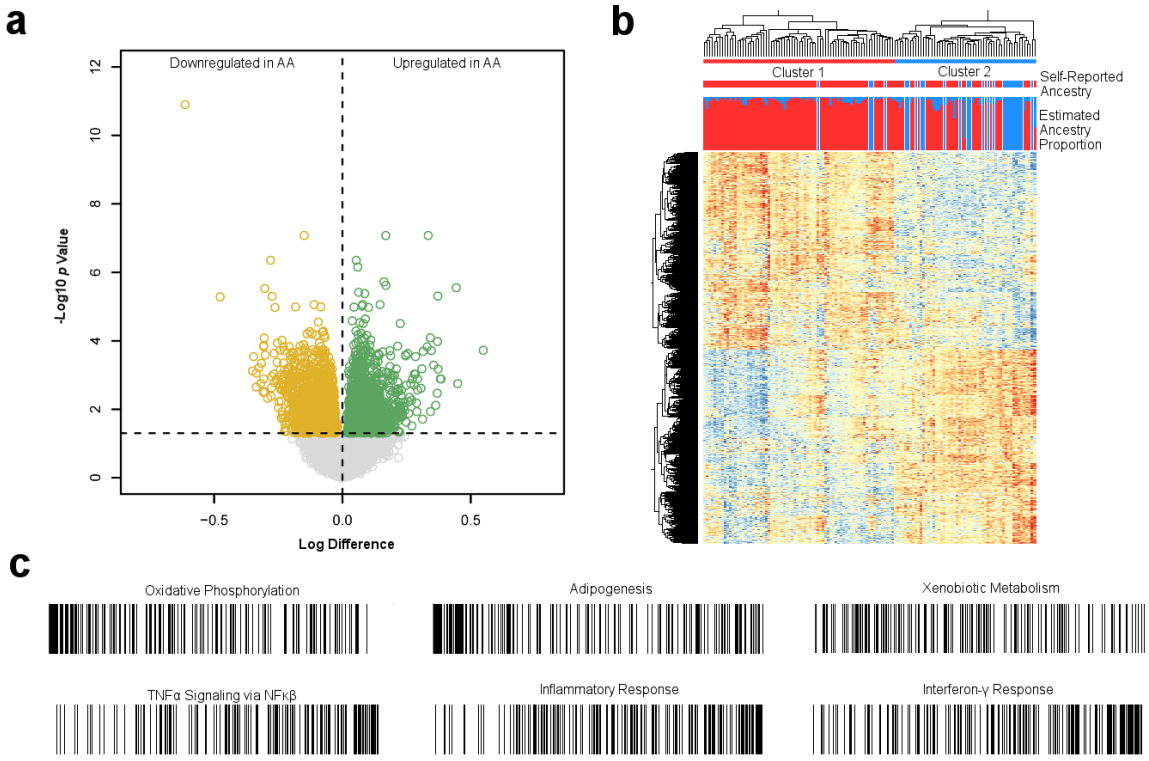


Figure 13 – Differential gene expression by ancestry. (a) Volcano plot depicting log difference (x-axis) and –log10 p-value (y-axis) for 14,392 genes between African (n=104) and European (n=25) ancestry IBD patients. (b) Hierarchical clustering of 2,705 genes differentially expressed at FDR < 0.05 from (a). Top bar: individuals grouped into Cluster 1 (red; n=70 AA, 4 European) and Cluster 2 (blue; n=34 AA, 21 European). Middle bar: self-reported AA (red) or European (blue) ancestry. Bottom bar: Estimated proportion of African (red) and European (blue) ancestry from supervised ADMIXTURE analysis. (c) Gene set enrichment analysis ranking plots. Each line represents a gene, while position from left to right represents ranking calculated by multiplying sign of fold change by inverse of FDR value. Top row shows bias towards downregulated genes in AA; bottom row shows bias towards upregulated genes.

Following differential expression analyses, we sought to validate that the population-based variation in gene expression we observed is heritable and not solely attributable to environmental differences. By studying an admixed population, we can contrast the influences of European and African ancestry among individuals of mixed ancestry to the difference between the two predominant ancestry groups. Proportions of admixture in AA were estimated from common genotypes called from the RNA-Seq reads, using ADMIXTURE software (163) with 1000 Genomes CEU and YRI individuals as reference European and African populations (Figure 14a). AA who group with European ancestry individuals in Cluster 2 exhibit a highly significant ($p = 6\times10^{-5}$) trend of increased proportions of European admixture compared with AA grouping in Cluster 1 (Figure 14b). We then applied methods introduced by Price et al. (164) to estimate the percentage of gene expression variation between populations that can be attributed to genetic effects. They argued that the slope of the regression of gene expression against ancestry proportion should be the same as the regression of the difference between the two predominant ancestry groups if the effect of ancestry is purely genetic, whereas if the two measures are uncorrelated it is purely environmental. The slope c of the two regressions assessed across all genes thus provides an estimate of the total sample heritability. By applying this method, we obtained c=0.43, which we validated as statistically significant ($p = 0.05$) relative to 1000-permutations (Figure 14c, Appendix A: Supplementary Table 7). This value implies a substantial heritable component to differences in gene expression observed between AA and European ancestry IBD patients.

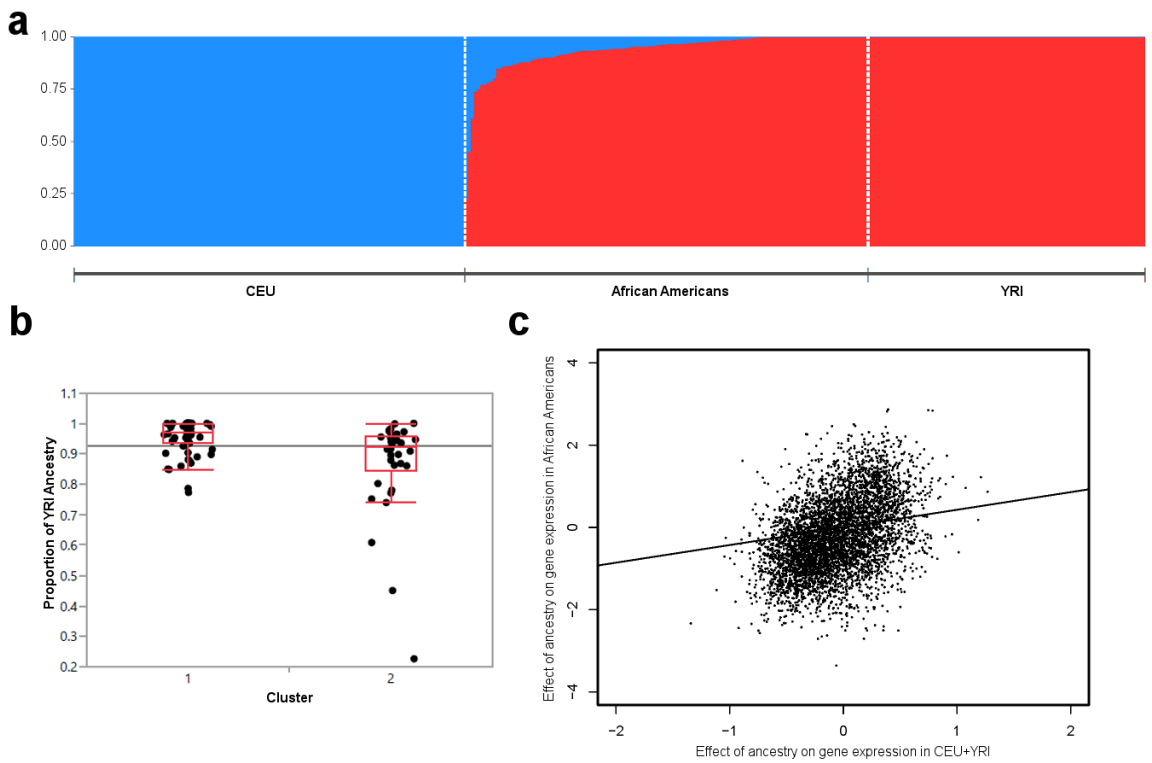


Figure 14 – Influence of ancestry proportions on gene expression. (a) Proportions of ancestry assigned to AA individuals. CEU population consists of 1000 Genomes Utah Northern and Western European ancestry individuals (n=85) and European ancestry individuals from this study; YRI population consists of 1000 Genomes Yoruban individuals (n=88). (b) Representative boxplot for ANOVA between self-identified AA in Cluster 1 (n=70) and Cluster 2 (n=34) from Figure 1b. Y-axis represents proportion of YRI ancestry. $P = 6\times10^{-5}$. Horizontal line is grand mean. (c) Plot of regressed estimates of the effect of ancestry proportion on gene expression for the top 5,000 most highly expressed genes, calculated in CEU vs YRI (x-axis) and within AA (y-axis). The slope of the correlation line is c=0.43.

3.4 Conclusions

In summary, our study shows strong differential gene expression in key pathways based on African versus European ancestry. We further demonstrate that this variation in gene expression amongst populations can be partially attributed to heritable, genetic effects, rather than solely to differences in environmental factors. The pathways highlighted here are known to be critical in IBD pathogenesis, and elevation of inflammatory and TNFα signaling appears to be consistent with evidence of worse prognosis in AA. Further investigation into ancestry-specific variation in disease is necessary for the development of personalized therapeutics.

CHAPTER 4. GENE EXPRESSION BASED STRATIFICATION OF RISK OF PROGRESSION TO COLECTOMY IN ULCERATIVE COLITIS

An important goal of clinical genomics is to be able to stratify risk of disease progression. Between 5% and 10% of ulcerative colitis (UC) patients require colectomy within five years of diagnosis (165), but polygenic risk scores (PRS) utilizing findings from GWAS are unable to provide meaningful prediction of this adverse status (166). By contrast, in Crohn's disease, gene expression profiling of GWAS-significant genes does provide some stratification of risk of progression to complicated disease in the form of a Transcriptional Risk Score (TRS) (8). Here we demonstrate that measured gene expression identifies UC patients at 5-fold elevated risk of colectomy with data from the PROTECT clinical trial (167). This chapter represents a collaborative effort of all of the members of the PROTECT consortium. I especially wish to recognize Drs. Jeffrey Hyams, Ted Denson, and Subra Kugathasan for their leadership efforts, and Kyle Gettler and Mamta Giri from the Cho lab at Mt. Sinai for their contributions to replication and single cell sequencing analyses.

4.1 Introduction

PROTECT is a multicenter pediatric inception cohort study of response to standardized colitis therapy (167). We have previously shown that a signature of rectal

mucosal gene expression at diagnosis, prior to therapeutic intervention, associates with corticosteroid-free remission with mesalamine alone observed in 38% of 400 patients by week 52 of follow-up (167). A signature of rectal mucosal gene expression associated with week 4 corticosteroid response in PROTECT is related to one indicative of response to anti-TNFα and anti-α4β7 integrin therapy in adults (168), and reciprocally, active pediatric UC was associated with suppression of mitochondrial gene expression, and increasing disease severity with elevated innate immune function. In order to more explicitly model progression to colectomy observed in 6% (25 of 400) of the patients within one year of diagnosis, we performed differential expression analysis between baseline rectal RNA-Seq biopsies of 21 patients who progressed to colectomy, and 310 who did not.

4.2 Methods

4.2.1 The PROTECT cohort

428 participants aged 4 to 17 years were enrolled from 29 centers across North America into the PROTECT study upon clinical, histological, and endoscopic diagnosis of ulcerative colitis. Patients with disease extent beyond the rectum, a Pediatric Ulcerative Colitis Activity Index (PUCAI) score of ≥ 10, no prior therapy for colitis, and negative enteric bacterial stool culture were eligible to participate. All baseline assessments and sample collections were performed prior to the initiation of therapy. Initial treatment with mesalamine, oral corticosteroids, or intravenous corticosteroids was decided based on mild, moderate, or severe PUCAI. Following the baseline assessment, follow-up

assessments were performed at 4, 12, and 52 weeks, with other therapeutic interventions administered based on guidelines for need for additional medical therapy. The study parameters are described in further detail in Hyams et al (169).

4.2.2 RNAseq data processing and differential expression analyses

RNA was isolated from 340 rectal biopsies taken at baseline and 92 rectal biopsies taken at week 52 follow-up. RNAseq was performed with the Lexogen QuantSeq 3' platform. Using FastQC, the single end 150 bp reads were trimmed and adapters were removed (156). Reads were mapped to human genome hg19 using hisat2, and the aligned reads were converted into read counts per gene with SAMtools and HTSeq in the default union mode (121, 122, 157, 170). The raw read counts were normalized via trimmed mean of M-values normalization with the edgeR R package (123).

Expression of the sex-specific genes *RPS4Y1, EIF1AY, DDX3Y, KDM5D*, and *XIST* was used to validate the gender of each individual, resulting in the removal of two mismatches. Further adjustment and removal of batch effects was performed with surrogate variable analysis (SVA) combined with supervised normalization (SNM) (124, 125). Race, gender, initial treatment group, time of sampling, and week 52 colectomy status were modeled with the SVA R package, where initial treatment group, time of sampling, and week 52 colectomy status were protected variables, which resulted in the identification of 28 confounding factors. Of these, five variables significantly correlated with protected variables were preserved, while the remaining 23 were statistically removed with SNM.

Two individuals that were outliers in a principal component analysis of total gene expression were removed.

Differential gene expression testing was performed based on colectomy status with the voom R package. Log fold change and Benjamini-Hochberg adjusted p-values were obtained for all genes. The first principal component of the top 150 genes differentially expressed at baseline between patients who required colectomy by week 52 follow-up (n= 21) and patients who did not (n= 310) formed the gene expression-based risk score for colectomy ($PC1_{col}$). This score is moderately correlated (r=0.46) with PC1 of overall expression of genes differentiating UC cases and controls, reported by Haberman et al (168).

Cross validation for $PC1_{col}$ was performed by randomizing colectomy status amongst individuals prior to differential gene expression testing and calculation of $PC1_{colRand}$, as in the calculation for $PC1_{col}$. ANOVA was performed between randomized colectomy and non-colectomy individuals, with results from 1000 such tests reported in Appendix B: Supplementary Figure 7.

We compared expression of the genes comprising $PC1_{col}$ at baseline and week 52 with Mayo score as a marker for mucosal healing (Appendix B: Supplementary Figure 8). $PC1_{col}$ was calculated as previously described in the subset of individuals with baseline gene expression. Additionally, a restricted $PC1_{col\text{-}wk52}$ was calculated by finding PC1 of the 150 genes used in the calculation of $PC1_{col}$, within the subset of individuals with week 52 gene expression. Change in PC1 score was simply calculated as the difference between

$PC1_{col}$ and $PC1_{col\text{-}wk52}$. All p-values were generated with analysis of variance (ANOVA) tests.

Transcriptional Risk Scores (TRS), first introduced by Marigorta et al. (8) for discriminating IBD cases versus controls, capture the summation of polarized expression of genes incorporated based on both proximity to IBD GWAS hits and presence of eQTL in peripheral blood. We generated the TRS with four different strategies, all of which gave similar highly significant differentiation between colectomy and no colectomy samples (Fig. S1). Model 1 was a GLM using the top 9 genes *RGS14, APEH, MRPL20, POP7, CDC42SE2, RORC, EDN3, PTK2B*, and *STAT3* that differentiate patients by colectomy status ($p < 0.1$), essentially the sum of the z-scores weighted by their magnitude of differential expression. Model 2 was a GLM using the 10 genes discussed in the text due to strong co-regulation and association with colectomy. Models 3 and 4 were based on all 26 genes, generated with a weighted GLM or simple PC1 score, respectively. All four scores are highly correlated, $r>0.8$, indicating that they are capturing similar aspects of differential expression (Appendix B: Supplementary Figure 12). We report Model 4 in the text. This TRS is highly correlated with $PC1_{col}$ ($r=0.64$).

Relative proportions of epithelial and immune contributions to total rectal gene expression reported in Appendix B: Supplementary Figure 9 were evaluated by computing PC1 of the expression of 200 genes upregulated specifically in the total epithelial or immune components of the single cell gene expression dataset reported by Smillie et al (50). We checked each PC to ensure that positive values associate with elevated expression of the respective genes, and compared the values at Baseline and Week 52.

4.2.3 Replication of colectomy risk score and cell-type enrichment

Surgical specimens from 210 ulcerative colitis patients undergoing bowel resection for IBD at Mount Sinai Health System and affiliated clinicians were recruited to be part of the Mount Sinai Crohn's and Colitis Registry (MSCCR) between December, 2013 and September, 2016 as described (171-173). The protocol required written informed consent that was approved by the Icahn School of Medicine at Mount Sinai Institutional Review Board (HSM#14-00210). Patients who were enrolled in the study were asked to provide blood and/or biopsies, which were collected during a colonoscopy planned for regular care. Clinical and demographic information was obtained through a questionnaire. Patients were treated with a range or medications, including corticosteroids, infliximab, azathioprine, and mesalamine. All macroscopically moderate-to-severely inflamed tissues were confirmed as active colitis by pathology examination provided by the Mount Sinai Hospital (MSH) Pathology Department. Freshly collected representative 0.5-cm-wide tissue fragments were isolated from surgical specimen samples, flash frozen, and stored at −80 °C.

RNA was isolated from frozen tissue using Qiagen QIAsymphony RNA Kit (cat.# 931636) and samples with RIN scores >7 were retained. One microgram of total RNA depleted of ribosomal RNA using the Ribozero kit (Illumina Cat # MRZG12324) was used for the preparation of sequencing libraries using RNA Tru Seq Kits (Illumina (Cat # RS-122-2001-48). These were sequenced on the Illumina HiSeq 2500 platform using 100 bp paired end protocol. Base calling from Images and fluorescence intensities of the reads was done in situ on the HiSeq 2500 computer using Illumina software, aiming for 70,000 paired end reads per sample. Short reads were mapped to the GRCh37/hg19 assembly (UCSC

Genome Browser) with 2-pass STAR, and processed using RAPiD, which is a RNA-seq analysis framework developed and maintained by the Technology Development group at the Icahn Institute for Genomics and Multi-scale Biology. Detailed quality control metrics were generated using the RNASeQC package. Raw count data was pre-filtered to keep genes with CPM>0.5 for at least 3% of the samples. After filtering, count data was normalized via the weighted trimmed mean of M-values and further variance stabilized using a logarithmic transformation. Normalized counts were further transformed into normally distributed expression values via the voom-transformation using a model that included technical covariates (processing batch, RIN, exonic rate and ribosomal RNA rate), while accounting for the intra-patient correlation across regions.

We repeated the transcriptional risk assessment analysis in this external dataset after normalization for gender, age, exonic RNA ratio, and rRNA level expression levels, using the prcomp function in R with the 150 genes from the PROTECT $PC1_{col}$, or the 26 gene TRS. The R package ggplot2 was then used to plot the distribution of PC1 for patients who did (10 patients) or did not (201 patients) have follow-up colectomies (Appendix B: Supplementary Figure 10). Additionally, we performed hierarchical clustering of single-cell gene expression data to identify cell types implicated by both the PC1 and TRS gene sets. Cell types enriched for PC1 genes included plasmacytoid dendritic cells, endothelial cells, group I innate lymphoid cells, fibroblasts, and macrophages.

4.2.4 SNP data processing and eQTL studies

The Affymetrix UK BioBank Axiom Array was used to perform genotyping of 424 individuals across 800,000 SNPs. Imputation was performed using IMPUTE2 software (132), after which quality control performed using PLINK was used to remove SNPs not in Hardy-Weinberg equilibrium at $p < 10^{-3}$, SNPs with a minor allele frequency < 1%, or a rate of missing data across individuals > 5% (130). Approximately 7 million imputed SNPs passed these thresholds and were tested in the eQTL analysis. SNPs within 250 kb of the start and stop sites of a gene were considered to be cis to the gene and tested for a potential eQTL association. Mapping was performed with the mixed linear modelling method in GEMMA, which tested a set of approximately 12 million SNP-gene pairs for associations at a common p-value threshold of 1×10^{-5} (133). Two separate comparative analyses were performed, where the initial set of eQTL mapping was performed on all 330 baseline samples and 87 week 52 follow-up samples, and the secondary analysis was performed on 78 matched samples only, where the same individual was profiled at both time points. The initial full analysis yielded 91,774 significant SNP-gene associations at baseline and 19,371 associations at week 52 follow-up, and the secondary matched analysis yielded 14,272 significant unique SNP-gene associations at baseline and 12,617 significant associations at week 52 follow-up. These were further refined to 1,317, 218, 186, and 166 peak SNP to unique gene associations, respectively.

4.2.5 Single cell sequence analysis of the lamina propria

The following single cell analyses were performed primarily by Kyle Gettler and Mamta Giri of the Cho lab at Mt. Sinai, and reviewed by myself for inclusion in this study. For the analyses reported in Appendix B: Supplementary Figure 11, we analyzed a total of 34,157 cells from paired inflamed rectum (n = 4) and uninflamed sigmoid colon (n = 5) from 4 UC patients undergoing treatment at Mount Sinai Hospital. Resected tissue biopsies were collected in ice cold RPMI 1640 (Corning Inc.) and processed within one hour after termination of the surgery. To limit biased enrichment of specific cell populations related to local variations in the intestinal micro-organization, we pooled twenty mucosal biopsies sampled all along the resected specimens using a biopsy forceps (EndoChoice). Epithelial cells were dissociated by incubating the biopsies in a dissociation medium (HBSS w/o Ca2+ or Mg2+ (Life Technologies) with HEPES 10mM (Life Technologies) and enriched with 5mM EDTA (Life Technologies)) at 37°C with 100 rpm agitation for two cycles of 15 min. After each cycle, the biopsies were vortexed vigorously for 30 seconds, and washed in complete RPMI media equilibrated at RT. They were transferred to digestion medium (HBSS with Ca2+ Mg2+, FCS 2%, DNase I 0.5mg/mL (Sigma-Aldrich) and collagenase IV 0.5mg/mL (Sigma-Aldrich)) for 40 min at 37°C with 100 rpm agitation. After digestion, the cell suspension was filtered through a 70mm cell strainer, washed in DBPS / 2% FCS / 1mM EDTA and spun down at 400 g for 10 min. After red blood cell lysis (BioLegend), dead cells were depleted using the dead cell depletion kit (Miltenyi Biotec, Germany), following manufacturer's recommendations. Viability of the final cell suspension was

calculated using a Cellometer Auto 2000 (Nexcelom Biosciences) with AO/PI dye. The exclusion was routinely 70% or higher live cell rate.

Single cells were processed through the 10X Chromium platform using the Chromium Single Cell 3′ Library and Gel Bead Kit v2 (10X Genomics, PN-120237) and the Chromium Single Cell A Chip Kit (10X Genomics, PN-120236) as per the manufacturer's protocol. In brief, 10,000 cells from single cell suspension were added to each lane of the 10X chip. The cells were partitioned into gel beads in emulsion in the Chromium instrument, in which cell lysis and bar-coded reverse transcription of RNA occurred, followed by amplification, fragmentation and 5′ adaptor and sample index attachment. Libraries were sequenced on an Illumina NextSeq 500.

We aligned reads to the GRCh38 reference using the Cell Ranger v.2.1.0 Single-Cell Software Suite from 10X Genomics. The unfiltered raw matrices were imported into R Studio as a Seurat object (Seurat v3.0.1 (174)). Genes expressed in fewer than three cells in a sample were excluded, as were cells that expressed fewer than 500 genes and with UMI count less than 500 or greater than 60,000. We normalized by dividing the UMI count per gene by the total UMI count in the corresponding cell and log-transforming. The Seurat integrated model (174) was used to generate a combined ulcerative colitis model with cells from both inflamed and uninflamed samples retaining their group identity. We performed unsupervised clustering with shared nearest-neighbor graph-based clustering, using from 1 to 15 principal components of the highly variable genes; the resolution parameter to determine the resulting number of clusters was also tuned accordingly. Cell types were assigned using known markers previously described for Crohn's disease (175).

Visualization of relative abundance of specific genes in each cell type was performed using Seurat functions in conjunction with the ggplot2 (176).

4.3 Results

The volcano plot in Fig. 15a shows down-regulation of 783 transcripts in the colectomy cases (red), and up-regulation of 1,405 transcripts (blue) at the experiment-wide threshold of $p < 4\times10^{-6}$. Gene set enrichment analysis (160) summarized in Fig. 15b highlights engagement of multiple pathways previously implicated in adverse outcomes in inflammatory bowel disease, including TNF and interferon signaling, and various signatures of inflammation and immune response (7, 177).

The first principal component ($PC1_{col}$) of the top 150 of these differentially expressed genes has a weak negative correlation with our previously reported signature of remission detected in a subset of 206 patients using a different RNA-seq protocol (168). With very high significance, it distinguishes the colectomy cases from non-progressors, as all but one case have PC1 scores greater than 10, a value exceeded by only 20 of the 317 non-colectomy cases (Fig. 15c). A 1000-fold cross validation test confirmed the significance of the $PC1_{col}$ predictor is several orders of magnitude greater than would be expected by chance (Appendix B: Supplementary Figure 7). All of the high $PC1_{col}$ individuals were placed initially on corticosteroids, the majority intravenously (Fig. 15d); the score also correlates with a gradient of disease severity indicated by baseline PUCAI (pediatric ulcerative colitis activity index) (178) and initial treatment. We also obtained rectal biopsy

RNA-seq data for 92 patients at week 52 and observed significant depression of the score (Fig. 15e), indicative of mucosal healing even in the cases with elevated initial gene activity (none of the follow-up cases were colectomy, since the surgical procedure had been performed earlier). Appendix B: Supplementary Figure 8 shows that PC1 remains associated with Mayo score even at week 52, and that the change in PC1 molecular score over time correlates with the degree of mucosal healing.

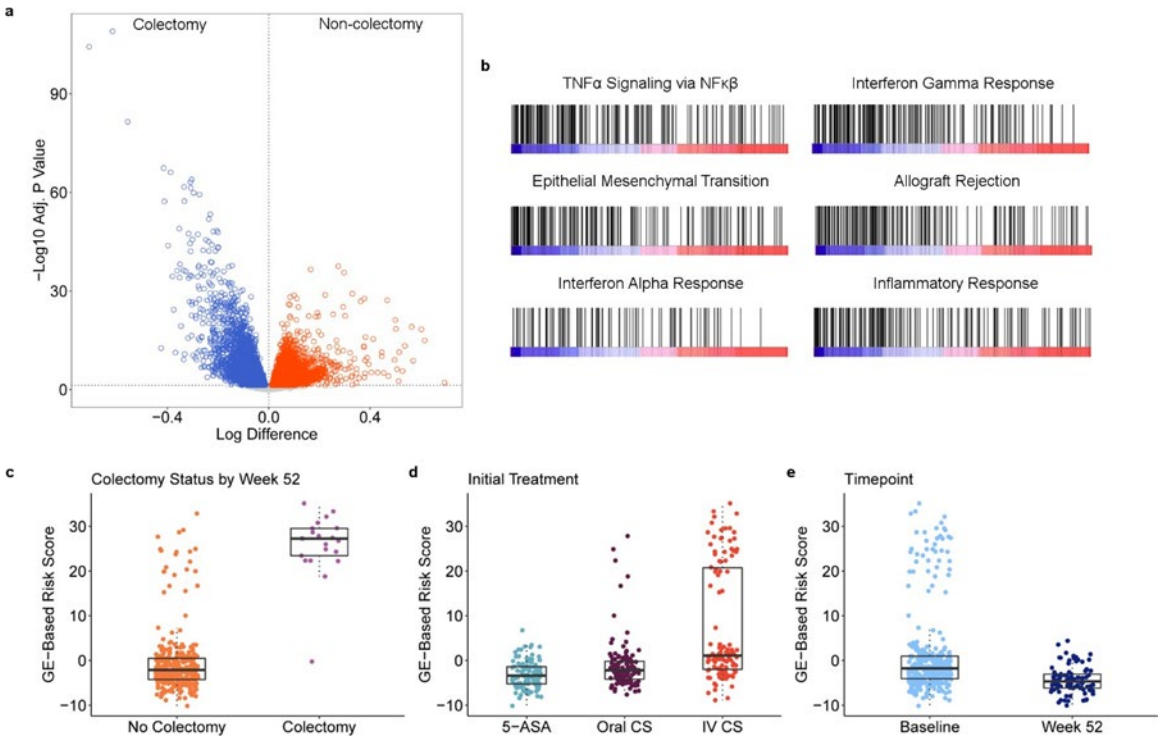


Figure 15 – Differential Expression Associated with Colectomy in the PROTECT study. (a) Volcano plot of significance (negative log10 of the p-value) against difference in expression on log2 scale, with genes up- regulated in colectomy in blue. (b) Six pathways highlighted by gene set enrichment analysis as up- regulated in colectomy. Each bar represents a gene in the indicated pathway, and position along the axis is representative of rank order of differential expression. From left to right, FDR = 0, 0, 0, 0, 2.43×10^{-4}, and 2.02×10^{-4}. PC1 of the differentially expressed genes as a function of (c) colectomy status at week 52; p = 2×10-45, (d) initial treatment; p = 5×10-20, and (e) baseline or week 52 follow-up biopsy profile; p = 2×10-7. All boxplots indicate 1st and 3rd quartile as box ends, with center median line and whiskers extending to farthest point within 1.5 times the interquartile range.

Given the marked shift in gene expression at follow-up, we next asked whether local regulation of the gene expression might contribute, by performing comparative eQTL analysis. Figure 16a indicates generally high concordance in the effect sizes (betas) at both time-points, with slight inflation of the estimates at baseline (1,416 blue effects) or week

52 (421 magenta effects), likely due to winner's curse. There were 72 eSNPs significantly regulating 308 genes at both time points, with the smaller number of eQTL at week 52 attributable to the smaller sample size. One quarter of the baseline eQTL are at least 2-fold greater than at week 52, and one third of the follow-up eQTL are at least 2-fold greater than at baseline. Clearly visible in Fig 16a are 33 apparently week 52-specific effects that are more than 20-fold greater than at baseline, the majority with reduced expression of the minor allele. Examples of baseline and follow-up specific eQTL affecting a variety of gene functions in immunity and epithelial cell biology are shown in Fig. 16b. Some of the change in eQTL profiles is likely attributable to an increase in the proportion of epithelial relative to immune cells at week 52 (Appendix B: Supplementary Figure 9).

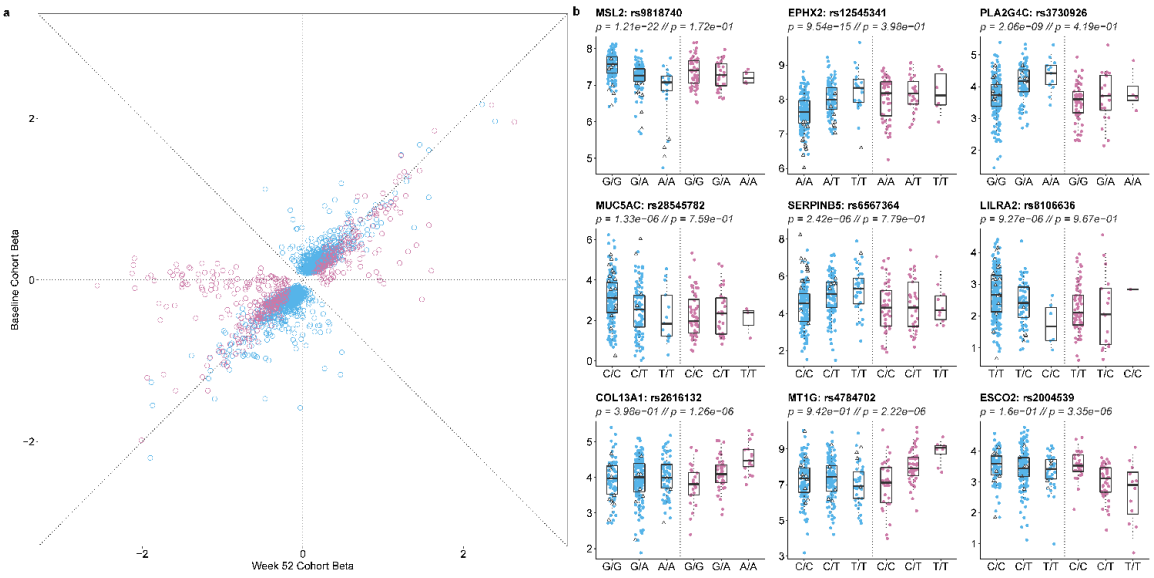


Figure 16 – eQTL contrast between baseline and week 52 follow-up in the PROTECT study. (a) Comparison of effect sizes (betas) for the effect of the minor allele on gene expression. Blue eQTL were discovered at baseline, and magenta only at week 52. (b) Examples of nine genes with differential eQTL effects at the two timepoints showing observed transcript abundance as a function of genotype at baseline or week 52 follow-up. The bottom row are genes with eQTL only at follow-up. All boxplots indicate 1st and 3rd quartile as box ends, with center median line and whiskers extending to farthest point within 1.5 times the interquartile range. Note that many of the genes with large negative follow-up betas in panel (a) have relatively small minor allele frequencies, hence insufficient homozygous minor allele genotypes to plot.

Next, we asked whether the intersection of GWAS, eQTL and differential expression could be used to generate a transcriptional risk score (TRS) for colectomy, analogous to the one we recently developed for prediction of risk of progression to complicated Crohn's disease (8). The heatmap in Fig. 17a showing the abundance of 26 transcripts included in the TRS_{IBD} derived with coloc overlap (137) of IBD GWAS and

peripheral blood eQTL signals, indicates striking enrichment for elevated or reduced expression of a dozen transcripts in the baseline rectal biopsies of PROTECT patients destined for colectomy. The strongest clusters include *RGS14, MRPL20, PTK2B, TNFRSF4, TNFRSF18* and *CDC42SE2* up-regulation, and *CISD1, EDN3, RORC*, and *PLA2R1* down-regulation. PC1 of the entire set of 26 genes results in a TRS_{UC} that discriminates colectomy from non-progressors at $p=1\times10^{-28}$ (Fig. 17b). A score above 3.24 has a sensitivity of 90% and specificity of 95% (Fig. 17c), generating a positive predictive value of 55%, which is nine times the prevalence of the rate of progression in the study. Corresponding likelihood ratios for positive and negative prediction are 18 and 10 respectively. TRS_{UC} also performs as well as the composite PC1 of all 2,500 differentially expressed genes.

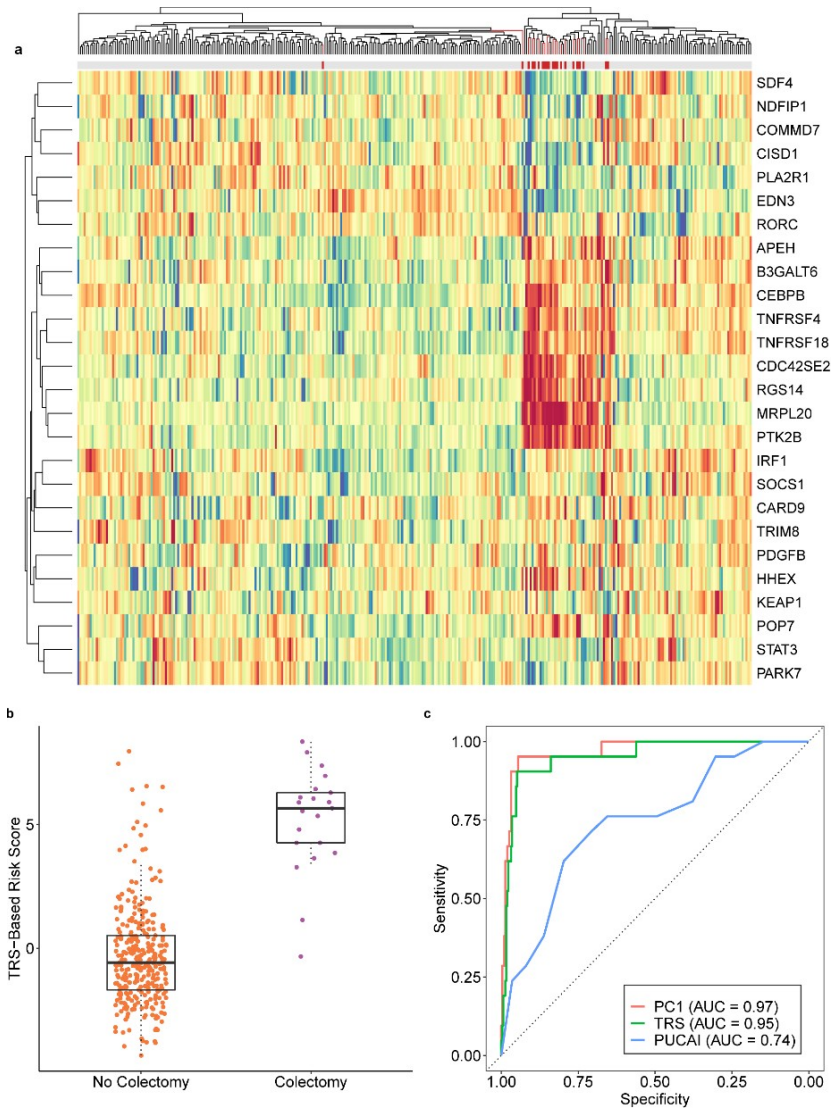


Figure 17 – Development of a Transcriptional Risk Score for Colectomy. (a) Heatmap of baseline rectal expression of 26 genes with evidence that the GWAS peak is the same as a blood eQTL (coloc H4 > 0.8), red high expression and blue low. The gray bar at the top indicates colectomy status, highlighting a cluster of patients for whom most of the genes are differentially expressed in the cases (red bars). (b) PC1 of the genes generates a TRS that is highly discriminatory between colectomy and non-colectomy at baseline; $p=1\times10^{-28}$. Boxplots indicate 1st and 3rd quartile as box ends, with center median line and whiskers extending to farthest point within 1.5 times the interquartile range. (c) Receiver operating characteristic curve contrasting sensitivity and specificity for colectomy showing that both the TRS (green) and PC1 of all differentially expressed genes (red) have high accuracy (AUC > 0.95), compared with PUCAI, a commonly used clinical disease severity index.

We replicated these findings in an independent adult ulcerative colitis cohort from Mt. Sinai Medical School in New York (172, 179). PC1 of the rectal expression of 146 genes strongly correlated with the PROTECT $PC1_{col}$ signature highly significantly ($p=0.0015$) distinguished 10 patients who have had colectomy from the remaining 201 (Appendix B: Supplementary Figure 10a), with the majority of genes differentially expressed in the same direction. Similarly, a TRS derived from the 26 GWAS-associated transcripts showed a strong trend toward differentiation of colectomy cases in the adult cohort (Appendix B: Supplementary Figure 10b), which was also highly significant ($p=0.010$) after removal of two outliers characterized by aberrant expression of *CDC42SE2*, the only transcript of the 26 tested which disagreed in direction of effect between the two studies.

Examination of the expression of colectomy-associated genes in a single cell RNA-seq dataset obtained from rectal biopsies provides strong evidence that both epithelial and immune cells contribute to the risk of disease progression (Appendix B: Supplementary Figure 11). Most of the genes are strongly expressed in just one or two of the 22 identified cell types, seven of which are notable for an excess of colectomy associated genes: plasmocytoid dendritic cells, immunoregulatory T-cells, ILC3 innate immune cells, and inflammatory macrophages from the immune compartment, and fibroblasts, secretory epithelial, and endothelial cells from the gut itself. The correlated expression of these gene sets suggests that risk of colectomy may in part reflect abnormal relative abundance of these cell types. On the other hand, each of these cell types is also represented in the single cell profiles of the TRS genes, which were selected on the basis of joint eQTL and GWAS

associations and hence are likely to be related to pathology through cis- regulatory effects. Prospective scRNAseq studies will likely reveal more insight into the cellular and genetic basis of the transcriptional risk of adverse disease progression.

4.4 Conclusions

Our results highlight the potential of transcriptional profiling for prediction of colectomy in ulcerative colitis. Direct measurement of rectal biopsy RNA provides a highly discriminatory signature observed in almost all children who will need surgery, and which predicts the adverse outcome in up to half of all cases. This expression profile reverts to a healthier state regardless of immunological therapy within one year. Our results are limited by the relatively small sample size of colectomies in the PROTECT study, which is nevertheless the largest treatment-naïve inception cohort to date. It is likely that more widespread sampling of this and other forms of inflammatory bowel disease will yield even more accurate predictors of disease progression, influencing personalized therapeutic decisions.

CHAPTER 5. SINGLE-CELL CHARACTERIZATION OF ILEAL EPITHELIAL CELLS IN CROHN'S DISEASE

5.1 Introduction

The gastrointestinal tract is a unique system of organs responsible not only for essential digestive and metabolic functions, but also for maintaining a highly delicate state of immune homeostasis (180). A perturbation of the balance between immune tolerance and response can result in the development of intestinal inflammation that is characteristic of inflammatory bowel disease.

The unique features of the intestinal epithelial cells that are primarily involved in this inflammation remain poorly characterized. Traditional bulk RNA-Seq experiments assume that tissues being sampled are comprised of homogeneous populations of cells. However, the intestinal epithelium is composed of highly heterogeneous cell types, including rare cell types and cells in varying states of development. Hence, bulk studies offer only a broadly averaged snapshot of gene expression across many different cells. In contrast, newer single cell sequencing technologies enable the dissection of cell-type specific contributions to gene expression in disease (181).

Several notable studies have sought to characterize gut mucosa at the single cell level (182). One of the earliest studies by Haber et al. in 2017 profiled intestinal epithelial cells obtained from mice, which enabled the initial identification of previously undescribed

subtypes of cells as well as gene markers of those populations (183). Numerous human-based studies soon followed, characterizing mucosal biopsies of the colon, ileum, and immune cells in the context of Crohn's disease and ulcerative colitis as well as in healthy individuals (49, 50, 175, 184-188).

Table 2 – Current single cell RNA-Seq studies of the human intestine in IBD

Year	Author	Biopsy Type	Focal Cells	Disease	Sample Size
2020	Elmentaite et al. (182)	Fetal gut, ileum	Epithelial	Healthy, CD	17 individuals - 62,854 cells
2020	Wang et al. (181)	Ileum, colon, rectum	Epithelial	Healthy	6 individuals - 14,537 cells
2019	Huang et al. (178)	Colon	Epithelial, stromal, immune	CD, UC	17 individuals - 73,165 cells
2019	Martin et al. (169)	Ileum	Stromal, immune	CD	11 individuals - 82,417 cells
2019	Parikh et al. (46)	Colon	Epithelial	Healthy, UC	6 individuals - 11,175 cells
2019	Smillie et al. (47)	Colon	Epithelial, stromal, immune	Healthy, UC	30 individuals - 360,650 cells
2019	Uniken Venema et al. (179)	Ileum	Immune	CD	3 individuals - 5,292 cells
2018	Kinchen et al. (180)	Colon	Stromal	Healthy, UC	10 individuals - 9,591 cells

The gut mucosa can be subdivided into three major cellular compartments—epithelial, stromal, and immune. In this chapter, I focus on the epithelial compartment of ileal biopsies obtained from healthy, treatment-naïve, and treated Crohn's disease patients. Intestinal epithelial cells serve as a barrier between the host and microorganisms coexisting

within the body, and act to modulate immune responses (189). Stem cells located at the base of the intestinal crypt work in conjunction with transit-amplifying (TA) cells to give rise to numerous differentiated cell lineages, including enterocytes, enteroendocrine cells, goblet cells, Paneth cells, tuft cells, and M cells. Of particular interest to us in this study are goblet cells, which secrete protective mucins. Prior non-single-cell studies have noted that degradation of goblet cell function contributes to the breakdown of the intestinal barrier in ulcerative colitis (190). A more recent scRNA-seq study by Parikh et al. uncovered diverse subpopulations of goblet cells within the colon and highlighted the existence of inflammation-associated subsets of goblet cells in ulcerative colitis (49).

Following up on these findings, in this chapter I examine ileal cell type proportions, gene expression, and associations with disease status in a cohort of 20 healthy, anti-TNFα naïve Crohn's disease, and treated Crohn's disease individuals. In particular, I examine the distinct subsets of goblet cells appearing in these ileal samples, which to my knowledge is the first such characterization specifically in Crohn's ileum. This study was performed in collaboration with the Qiu Lab at Georgia Tech and Kugathasan lab at Emory University as members of the Gut Cell Atlas Consortium, with funding awarded by the Helmsley Charitable Trust.

5.2 Methods

5.2.1 Cohort

A total of 6 healthy controls, 7 treatment-naïve Crohn's disease, and 7 Crohn's disease patients were profiled in this study. Of the 20 total participants in the study, 11 were of self-identified African American ethnicity, 7 were of Caucasian ethnicity, and 2 were of South Asian ethnicity. The majority of the cohort was considered to be pediatric, ranging in age from 9 years to 20 years old, with the exception of one adult patient aged 47 years. The group was split approximately evenly by gender, with 8 female participants and 12 male participants. Biopsies of the ileum were cryopreserved for single cell RNA sequencing, which was performed in three batches of 4, 8, and 8 samples with the 10x Genomics Chromium platform.

5.2.2 Single cell RNA-Seq processing

FASTQ files and alignment to GRCh38 was performed with Cell Ranger's mkfastq and count functions (191). Samples were then read into R using the Seurat package (174). A total of approximately 90,000 cells was analyzed across 20 samples. For the initial step of identifying epithelial cells, within the Seurat v3 framework, each of the samples was individually $\log_2$ normalized with a scale factor of 10,000 and scaled with the ScaleData function. The RunPCA function was used to identify PCs based on the top 2000 variable features. Finally, FindNeighbors, FindClusters, and RunUMAP were run to generate cluster assignments and UMAP clustering visualization based on 15 PCs.

Within individually normalized samples, clusters were assigned to one of three major cell type groupings—epithelial, stromal, and immune—based on highest average gene expression of the marker genes reported in Smillie et al. for epithelial cells (*EPCAM, KRT8, KRT18*), stromal cells (*COL1A1, COL1A2, COL6A1, COL6A2, VWF, PLVAP, CDH5, S100B*), and immune cells (*CD52, CD2, CD3D, CD3G, CD3E, CD79A, CD79B, CD14, CD16, CD68, CD83, CSF1R, FCER1G*) (50). A total of 68,241 cells grouping in clusters assigned as epithelial subsets were extracted from the raw counts data for the following joined sample set normalization steps.

5.2.3 Cell type annotation

Each of the raw epithelial-subset samples was then normalized with SCTransform individually. The standard recommended 2000 integration features were selected for running anchor integration with the IntegrateData function. Following integration, samples were again clustered as previously described. Assignment of cell types was based on gene markers reported in the healthy ileum by Wang et al. for enterocytes (*ALPI, SLC26A3, TMEM37, FABP2*), goblet cells (*ZG16, CLCA1, FFAR4, TFF3, SPINK4*), Paneth cells (*LYZ, CA7, SPIB, CA4, FKBP1A*), enteroendocrine cells (*CHGA, CHGB, CPE, NEUROD1, PYY*), progenitor cells (*SOX9, CDK6, MUC4, FABP6, PLA2G2A, LCN2*), transit-amplifying cells (*KI67, PCNA, TOP2A, CCNA2, MCM5*), stem cells (*LGR5, RGMB, SMOC2, ASCL2*), and tuft cells (*POU2F3, GFI1B, TRPM5*).

5.2.4 Gene expression analyses

Following assignment of cell types to clusters, differential expression analysis was performed between cluster 16 against all other clusters, and each individual sub-cluster of goblet cells against all other goblet cell clusters, using the FindMarkers function's implementation of the Wilcoxon rank sum test. Pathway annotation was performed with ToppFun, using genes upregulated and downregulated with an adjusted p-value of 0.05 or greater (192). All visualizations were generated with the ggplot2 package in conjunction with Seurat (176).

5.3 Results

5.3.1 Annotation of key epithelial cell subtypes

Increasingly precise classifications of cell types have become possible with the advent of single cell sequencing technologies. Despite the recent publication of several single cell transcriptomic surveys, many hurdles remain for the establishment of robust, inter-experimentally replicable cell type assignments. The fine resolution of single cell data demands equally precise reference datasets to enable accurate classification of cell types underpinning differential gene expression and other downstream analyses. At this nascent stage of development of the field, ideal reference datasets may only exist for some of the most commonly studied tissue types, such as healthy peripheral blood mononuclear cells.

In this chapter, I profiled human ileal epithelial cells biopsied from Crohn's disease patients, a very specific subset of cells which to date has been rarely profiled. Hence, I prioritized the matching of both organism and tissue type over disease status and selected the single-cell transcriptomic survey of the healthy human ileum performed by Wang et al. to serve as the reference gene expression set for our study. Wang et al. reported their comprehensive survey of about 15,000 epithelial cells from the ileum, rectum, and colon, building upon prior knowledge established in the mouse model organism (187). Here, following the extraction of epithelial cells and SCTransform-based normalization and integration described in the Methods, I utilized the gene markers they reported to annotate 28 Seurat-identified clusters to eight major cell types—enterocyte, enteroendocrine, goblet, Paneth, progenitor, stem, TA, and tuft.

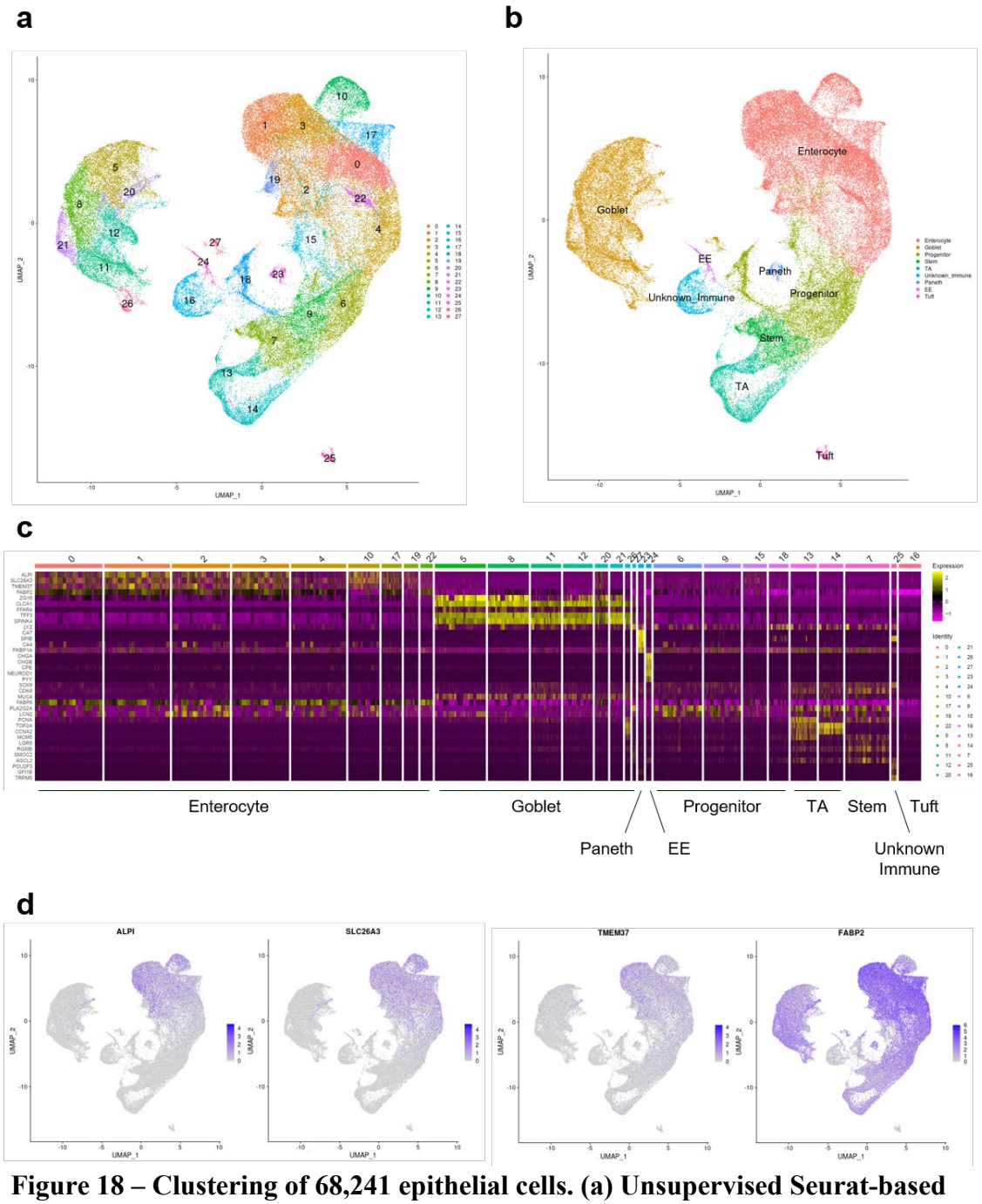


Figure 18 – Clustering of 68,241 epithelial cells. (a) Unsupervised Seurat-based clustering of 68,241 epithelial cells results in the identification of 28 clusters. (b) Cells were classified into 8 broad categories of cell types—enterocyte (n=31,329), progenitor (n=10,587), goblet (n=15,347), stem (n=3,619), TA (n=4,159), Paneth (n=485), enteroendocrine (n=444), tuft (n=435), and unknown immune (n=1,836). (c) Gene expression of Wang et al. markers across clusters utilized to assign cell types. (d) Representative gene markers of enterocytes. The left 3 genes serve to distinguish clusters well, while the rightmost shows broad expression throughout the dataset.

In their own analysis, Wang et al. based cell assignments upon previously published cell markers (183, 193). They reported that certain markers, for example, the previously reported tuft cell marker *DCLK1,* was not detected in their dataset. Similarly, in my analysis I found the expression of certain marker genes did not distinguish clusters well, either due to low expression or overly broad expression. Panel (d) of Figure 18 illustrates examples of 3 gene markers that performed well to distinguish a subset of cells, and 1 which performed poorly within our dataset. Additionally, one cluster, cluster 16, was ambiguously characterized based on the Wang et al. gene markers and may potentially contain a novel subset of cells unrepresented in the panel. Pathway analysis of genes differentiating cluster 16 against all other clusters within the dataset revealed upregulation of genes in pathways related to antigen binding and immune response, and downregulation of genes in oxidoreductase and mitochondrial activity. However, because these cells were filtered from the whole sample for stronger expression of epithelial genes over immune cell markers, it is unlikely that this cluster simply represents a cluster of typical immune cells. Hence, here I have labelled the ambiguously defined cluster as "Unknown Immune". Appendix A: Supplementary Table 8 lists top relevant pathways implicated by the differential expression analysis.

5.3.2 Disease status associations with differences in cell type proportions and extreme gene expression

Following the establishment of cell type annotations to clusters, I examined proportions of each of the sub-populations within each of the disease categories (healthy, treatment-naïve, and Crohn's disease). A few patterns of differential enrichment of cell

type clusters based on disease status emerge. Figure 19 visualizes these proportions across disease groups and cell types, while Table 3 lists these proportions.

Table 3 – Cell type proportions by disease status

Disease	Cell Type	Proportion of Disease Group in Cell Group	Proportion of Cell Group in Disease Group	Wang et al. Proportion
CD	Enterocyte	0.346	0.493	
Control	Enterocyte	0.261	0.479	~0.72
TN-CD	Enterocyte	0.393	0.422	
CD	Enteroendocrine	0.369	0.007	
Control	Enteroendocrine	0.306	0.008	~0.005
TN-CD	Enteroendocrine	0.324	0.005	
CD	Goblet	0.272	0.189	
Control	Goblet	0.229	0.206	~0.06
TN-CD	Goblet	0.500	0.263	
CD	Paneth	0.254	0.006	
Control	Paneth	0.384	0.011	~0.01
TN-CD	Paneth	0.363	0.006	
CD	Progenitor	0.342	0.164	
Control	Progenitor	0.224	0.139	~0.10
TN-CD	Progenitor	0.434	0.158	
CD	Stem	0.300	0.049	
Control	Stem	0.314	0.067	~0.05
TN-CD	Stem	0.386	0.048	
CD	TA	0.289	0.055	
Control	TA	0.284	0.069	~0.05
TN-CD	TA	0.426	0.061	
CD	Tuft	0.294	0.006	
Control	Tuft	0.389	0.010	Not reported
TN-CD	Tuft	0.317	0.005	
CD	Unknown Immune	0.364	0.030	
Control	Unknown Immune	0.107	0.011	Not reported
TN-CD	Unknown Immune	0.529	0.033	

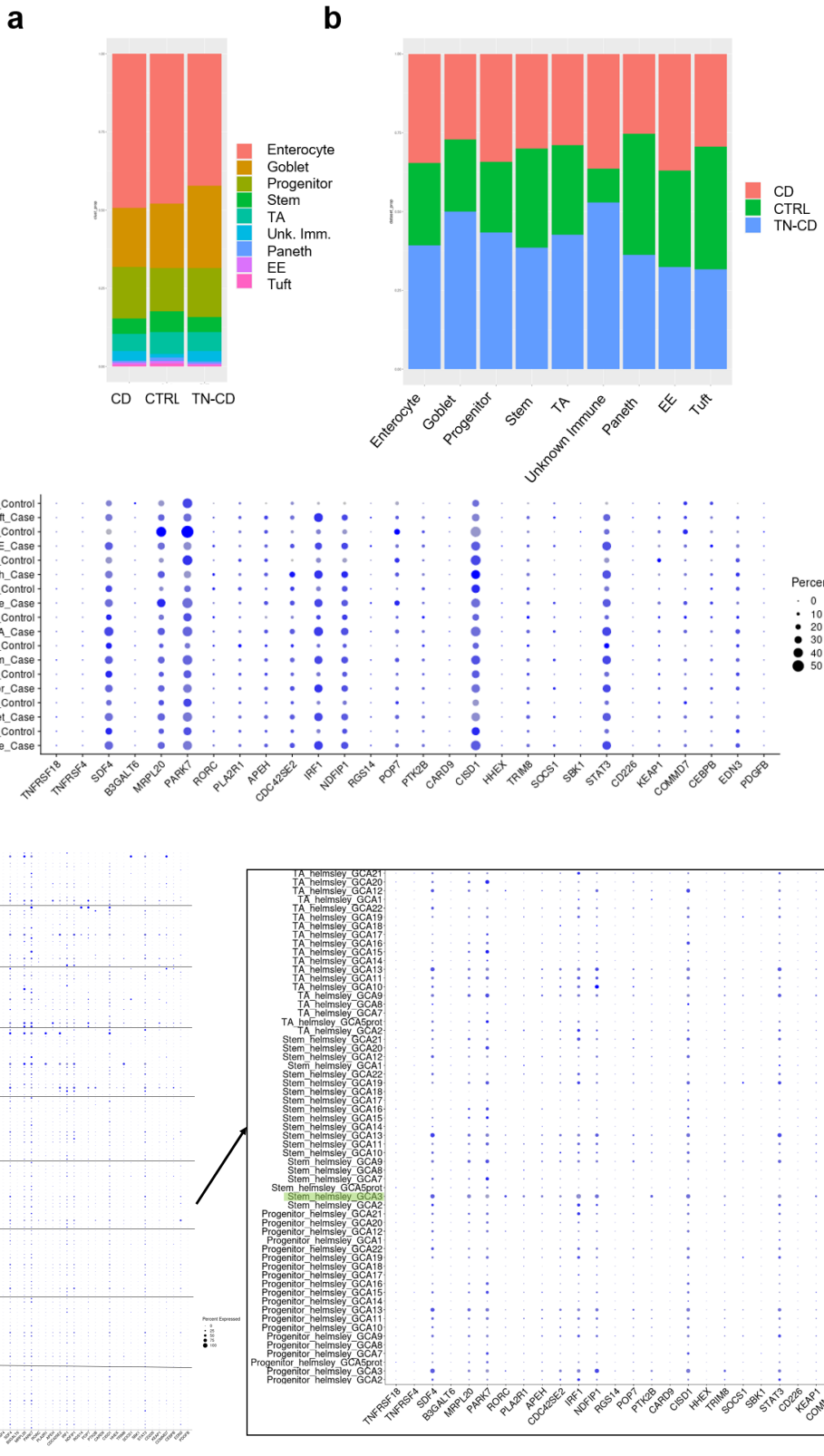

a
Enterocyte
Goblet
Progenitor
Stem
TA
Unk. Imm.
Paneth
EE
Tuft
CD
CTRL
TN-CD
b
CD
CTRL
TN-CD
Enterocyte
Goblet
Progenitor
Stem
TA
Unknown Immune
Paneth
EE
Tuft
c
Tuft_Control
Tuft_Case
EE_Control
EE_Case
Paneth_Control
Paneth_Case
Unknown_Immune_Control
Unknown_Immune_Case
TA_Control
TA_Case
Stem_Control
Stem_Case
Progenitor_Control
Progenitor_Case
Goblet_Control
Goblet_Case
Enterocyte_Control
Enterocyte_Case
TNFRSF18
TNFRSF4
SDF4
B3GALT6
MRPL20
PARK7
RORC
PLA2R1
APEH
CDC42SE2
IRF1
NDFIP1
RGS14
POP7
PTK2B
CARD9
CISD1
HHEX
TRIM8
SOCS1
SBK1
STAT3
CD226
KEAP1
COMMD7
CEBPB
EDN3
PDGFB
Percent Expressed
0
10
20
30
40
50
d
Tuft
EE
Paneth
Unknown Immune
TA
Stem
Progenitor
Goblet
Enterocyte

Figure 19 – Cell type proportions associated with disease status. (a) Proportions of cell types constituting samples grouped by control (n=17,080), Crohn's disease (n=22,002), and treatment-naïve Crohn's disease (n=29,159). (b) Proportions of each of the major groups of cell types originating from each of the disease status categories. (c) Dot plot visualizing expression of TRS genes (x-axis) in each of the major cell type groups subdivided by case (Crohn's disease and treatment-naïve Crohn's disease) and control. (d) Select expanded subset of dot plot visualizing expression of TRS genes (x-axis) in each of the major cell type groups subdivided by individual. An example of extreme expression in an individual's cell type is highlighted in green. See Appendix B: Supplementary Figure 13 for the full plot.

The proportions of most cell types identified in this study show reasonable concordance with previously reported data from the healthy ileum by Wang et al., with the exception of an increased proportion of goblet cells irrespective of disease status conversely also resulting in a reduced proportion of enterocytes. Notably, the Unknown Immune cluster appears to be largely composed of cells from the two Crohn's disease groups, with cells from the control samples constituting only about 10% of the cells. This, in conjunction with the immune-related pathways implicated by differential gene expression, suggest that this subset of cells may be related to response to disease. Additionally, there is some evidence of differences in proportions of goblet cells between the three disease states. The control and Crohn's disease samples contain fewer goblet cells than the treatment-naïve Crohn's samples, although the difference does not reach the threshold of significance with ANOVA testing (p = 0.13). Prior surveys of goblet cell proportions tended to focus on the colon rather than the ileum, although one review did

report that goblet cell proportions generally increase from the duodenum to distal colon, in association with increasing proportions of bacteria (194).

As an initial examination of the potential utility of Transcriptional Risk Scores in single cell data, I compared the expression of 28 TRS genes as described by Marigorta et al. within case-control and cell type groupings (8). Differences in gene expression by disease status are apparent, for example in *IRF1* ($p < 2.2\times10^{-308}$), *STAT3* ($p < 2.2\times10^{-308}$), and *NDFIP1* ($p = 2.3\times10^{-103}$). By further breaking down gene expression by cell type to sample, we observe instances of outlier individual-driven gene expression, for example in the high expression of *EDN3* in stem cells of Crohn's disease patient GCA3 (Figure 19d and Appendix B: Supplementary Figure 13). These suggest an alternate perspective of disease gene expression associations, in contrast to the broadly averaged case-control approach, where risk for disease may be mediated by different genes amongst individuals and thus may be more appropriately targeted with personalized therapeutics. Following on the example given here, we can imagine that amongst clinically similar Crohn's disease patients, the aberrant expression of *EDN3* in patient GCA3's stem cells might suggest a more successful drug target unique to their transcriptomic profile.

5.3.3 Distinct subtypes of goblet cells associated with disease status

Goblet cells are critical for proper functioning of the mucosal barrier, and abnormalities in mucus secretion are well known to be associated with the onset of intestinal inflammation in ulcerative colitis. However, as mentioned earlier, goblet cells in

the ileum, especially in the context of Crohn's disease, remain poorly characterized. A prior, non-single-cell-based survey of goblet cells in the colon found that proportions were moderately reduced in Crohn's disease and the presence of goblet cell differentiation factors was increased in inflamed Crohn's disease samples (190). In this study, following annotation of the nine broad categories of cell types, I further examined the subclusters of cells constituting the broad goblet cell cluster. I observed eight subclusters of goblet cells in our dataset, each differentiated by varying pathways implicated by gene expression. Table 4 summarizes these pathways. Defects in ribosomal synthesis have previously been linked with increased goblet cell differentiation, and upregulation of oxidoreductase activity may be indicative of stress responses to inflammation.

Table 4 – Differentially regulated pathways amongst goblet subclusters

Cluster	Upregulated	Downregulated
5	Adhesion & Defense Response	Ribosomal Activity
8	Intestinal Epithelium	Ribosomal Activity
11	Ribosomal Activity	Peptide Antigen Binding
12	Oxidoreductase Activity	Protein & TF Binding
20	Oxidoreductase Activity	Ribosomal Activity
21	Ribosomal Activity	Peptide Antigen Binding
26	Ribosomal Activity	Peptide Antigen Binding
27	Carbohydrate Binding	Proton Transporter Activity

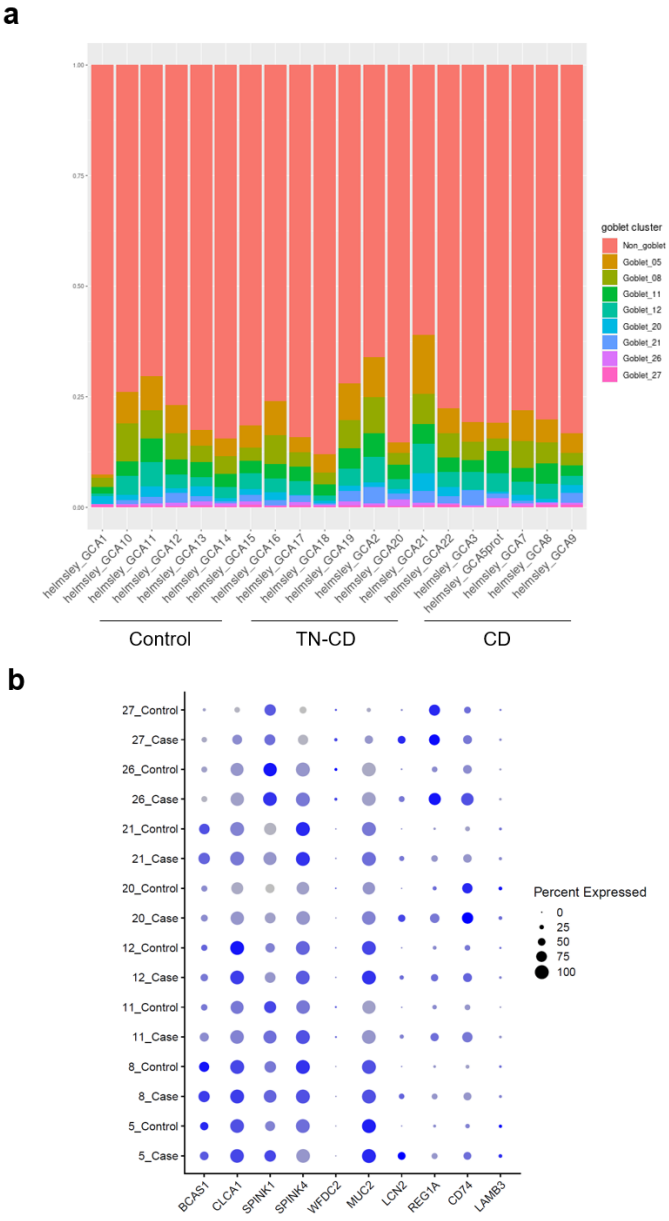


Figure 20 – Breakdown of differential proportions of goblet subclusters by individual and disease status. (a) The y-axis depicts proportion of cells constituting each sample (x-axis) subdivided by non-goblet and the eight clusters of goblet cells. (b) Expression of reported subcluster and spatial marker genes from Parikh et al. (49) subdivided by case-control status.

Parikh et al. previously reported in their single-cell study of the UC colon evidence of five subclusters of goblet cells with differential gene expression associations with disease (49). A number of gene markers of note that they report in their goblet cell analysis include *BCAS1* (cluster 5), *CLCA1* (crypt bottom, cluster 1), *SPINK1* and *SPINK4* (expressed in healthy crypt bottoms and inflamed crypt tops), *WFDC2* (crypt bottom, reduced expression in inflammation, cluster 2), *MUC2* (general marker of goblet health), *LCN2* and *REG1A* (expressed throughout crypt), *CD74* (crypt bottom), and *LAMB3* (crypt top). Figure 20b shows the expression of these genes in our dataset. Understandably, as their study was performed in UC colon as opposed to the CD ileum profiled in this chapter, several of the markers are too lowly or broadly expressed to be utilized for inferring mapping between clusters. A few interesting patterns can be observed here, such as expression of the crypt bottom marker *CD74* in cluster 20 ($p = 3.11\times10^{-200}$ for 20 vs. all other clusters) and non-disease-specific absence of the focal UC inflammation-associated gene *WFDC2* (expressed in approximately 2% of cells, p = 1 for case vs. control). Taken as a whole, it appears that goblet subclusters are not directly transferrable between studies and mechanisms for goblet cell associations with disease differ between UC and CD, a conclusion supported by previous non-scRNA-Seq based studies (190, 195).

5.4 Conclusions

In this study, I reported ileal cell type proportions, gene expression, and associations with disease status in a cohort of 20 healthy individuals and Crohn's disease patients. Additionally, I highlighted subclusters of goblet cells and potential associations with disease. One of the primary goals of this study was to establish a robust, standardized workflow for analysis of this unique single cell ileal epithelial dataset. To that end, several challenges and avenues of exploration remain to be addressed. Foremost, as was touched upon earlier, the inter-experimental replicability of certain gene markers, particularly for less commonly profiled cell types such as the ones we have analyzed, must be improved to enable appropriate cell type classification and accurate downstream gene expression analyses which are dependent on these assignments. As more and more single cell surveys are published, the increasing numbers of cells being profiled should ameliorate this issue in time. Another challenge to be addressed is the development of consistent bioinformatics algorithms for clustering that are just as important for accurate cell type assignment. Our group is now working on such strategies to improve the consistency of findings. One approach I am in the early stages of developing is a technique tentatively named "iterative clustering", which is based upon repeated clustering attempts to identify a core set of highly confidently grouped cells. As the field of single cell is still very much growing and developing, I expect that numerous other strategies for improving the replicability of analyses will emerge.

CHAPTER 6. CONCLUSIONS AND FUTURE DIRECTIONS

Extensive genome-wide association studies of inflammatory bowel disease have left us with a wealth of loci to investigate. Uncovering the links between these genetic risk loci and the underlying mechanisms of disease is the fundamental aim of this thesis. The potential benefits of genomic and transcriptomic directed personalized medicine approaches to IBD are especially great, as early therapeutic interventions have been shown to slow the progression of disease in patients who would have otherwise experienced severe complications.

The foundation upon which each of the analyses I have shared here is built is transcriptomic data. In chapter 2, I discuss a comparative analysis of juvenile idiopathic arthritis and inflammatory bowel disease which serves to highlight transcriptomic similarities and differences between two clinically heterogeneous and immune-mediated disorders. RNA-Seq data was also key to calculations performed in chapter 3 demonstrating the heritable differences in gene expression between individuals of varying proportions of African ancestry. Across chapters 2 and 4, I describe mapping of eQTL in several states—across diseases, and timepoints. The utility of gene expression data for investigating potential causal mechanisms underlying genetic variants is underscored here, as despite the relatively small sample size in comparison with typical GWAS studies, the greater impact of individual regulatory variants on gene expression enables the mapping of eQTL and comparison of effect sizes across conditions. Furthermore, I sought to expand upon the previously described TRS, which incorporates GWAS, eQTL, and transcriptomic

data, and demonstrated its ability to discriminate not just case-control status, but risk of progression to colectomy in ulcerative colitis. Finally, the advance to single cell RNA sequencing has enabled my most recent examinations of cell-type specific differences in gene expression associated with inflammatory bowel disease.

Several barriers remain to the implementation of genomics-based precision medicine in IBD, and beyond. A few such challenges of particular interest to me that fall within the scope of this thesis are the following—accounting for population and ancestry-specific differences, developing techniques for replicable analysis of single cell data, and building upon the transcriptional risk score and other potential gene expression based predictive measures.

It is well known that allele frequencies, linkage disequilibrium structure, and environmental exposures differ between populations(161, 196-198). With that in mind, it is rather unsurprising that studies have repeatedly demonstrated that genetic and polygenic risk scores based on European GWAS often perform poorly and do not reflect true risk of disease in populations not of European ancestry (199). From the perspective of evolutionary genomics as well, it makes sense that population-specific differences in immune response exist (63, 64). In order for genetics-based risk scores to be successfully and equitably applied, it is important for individual researchers to be cognizant of the potential for European-centered datasets to bias findings in other populations, and for the research community collectively to support increasing representation of populations of diverse ancestries.

With the development of the field of single cell transcriptomics, increasingly large datasets are being published and disseminated amongst researchers. Several micro-challenges fall under the umbrella of improving the inter-experimental replicability of single cell analyses. Starting with the data-driven, the establishment of consistent gene markers for assignment of cell types is critical for downstream analyses such as differential gene expression. I am optimistic that with time, the costs of single cell RNA-seq will decline as with older sequencing technologies, resulting in the establishment of cell atlases encompassing a wide spectrum of tissues and cell states. From the bioinformatics angle, algorithms which can reproducibly perform unsupervised clustering of cells are just as important for accurate identification of cell types. In addition, other challenges unique to single cell data, such as handling large amounts of dropouts, must be resolved. To begin to address this issue of replicability, I have begun to develop my own approach to identifying highly confidently clustered cells via iterative clustering attempts.

Finally, the transcriptional risk score represents a linkage between GWAS loci, eQTL, gene expression, and disease. In this thesis, I have reported my varied applications of TRS and its principles, as well as my own version of a predictive score for risk of disease complication in chapter 4. The transcriptional risk score could be further expanded on in any number of ways, but a future direction I would especially like to highlight is single cell TRS. I believe that the finer resolution of single cell gene expression data will further reveal potential mechanisms of association with disease, and possibly explain some of the incoherent associations initially reported in the original paper.

In summary, this thesis has demonstrated the applicability of genomic and transcriptomic profiling for characterization and prediction of risk of disease in inflammatory bowel disease. Many of the principles and strategies described here are also broadly applicable across diseases. Several hurdles remain to be overcome before such precision medicine approaches can be successfully applied, but they will offer incredible advantages in guiding therapeutic decision-making and ensuring that patients receive the most optimal treatment tailored to their individual needs, reducing the burden of healthcare on society and individuals.

APPENDIX A. SUPPLEMENTARY TABLES

Supplementary Table 1 – List of genes included in Transcriptional Risk Score.

A. RA-TRS (transcriptional risk score based on rheumatoid arthritis GWAS)

GWAS information (Okada et al. Nat Gen, 2014)									Target Gene	eQTL activity at GWAS SNP (eQTLbloodbrowser)					eQTL activity at SNP in LD with GWAS SNP					
GWAS_SNP	chr	poshg19	A1	A2	A1_FreqCase	A1_FreqCtrl	OR	Pvalue	Target Gene	GWAS_SNP_eQTL_effect	GWAS_SNP_eQTL_effect_all	GWAS_SNP_eQTL_effect_effall	GWAS_SNP_eQTL_effect_effdir_zscore	GWAS_SNP_eQTL_effect_pval	GWAS_SNP_inLD_eQTL_effect	GWAS_SNP_inLD_eQTL_all	GWAS_SNP_inLD_eQTL_effall	GWAS_SNP_inLD_eQTL_effdir_zscore	GWAS_SNP_inLD_eQTL_pval	Inferred Direction of Risk
rs28411352	1	38278579	T	C	0.27	0.24	1.12	4E-12	INPP5B	-	-	-	-	-	rs2306627	T/C	T	9.8	7E-23	High Expr
rs2476601	1	114377568	A	G	0.16	0.10	1.81	#####	PTPN22	rs2476601	G/A	A	6.3	3E-10	-	-	-	-	-	High Expr
rs2228145	1	154426970	A	C	0.62	0.60	1.08	4E-09	IL6R	-	-	-	-	-	rs4537545	T/C	T	-11.3	2E-29	High Expr
rs2317230	1	157674997	T	G	0.44	0.42	1.08	2E-08	FCRL3	-	-	-	-	-	rs2210913	C/T	T	35.6	#####	High Expr
rs72717009	1	161405053	T	C	0.11	0.10	1.13	2E-07	FCGR2B	-	-	-	-	-	rs7529225	G/A	A	21.8	#####	High Expr
rs34695944	2	61124850	T	C	0.69	0.73	0.89	3E-13	REL	-	-	-	-	-	rs13017599	G/A	A	3.8	1E-04	High Expr
rs1980422	2	204610396	T	C	0.79	0.81	0.89	6E-12	CD28	rs1980422	T/C	C	-8.6	8E-18	-	-	-	-	-	Low Expr

rs73081554	3	58302935	T	C	0.08	0.07	1.18	5E-08	PDHB	-	-	-	-	-	rs6772228	T/A	A	-11.6	4E-31	Low Expr
rs2561477	5	102608924	A	G	0.29	0.31	0.92	2E-09	PAM	rs2561477	G/A	A	22.6	#####	-	-	-	-	-	Low Expr
rs657075	5	131430118	A	G	0.17	0.17	1.09	1E-06	ACSL6	rs657075	G/A	A	6.9	4E-12	-	-	-	-	-	High Expr
rs2451258	6	159506600	T	C	0.75	0.74	1.11	2E-10	RSPH3	rs2451258	T/C	C	-4.0	7E-05	-	-	-	-	-	High Expr
rs1571878	6	167540842	T	C	0.51	0.55	0.86	6E-30	RNASET2	rs1571878	C/T	C	-37.2	#####	-	-	-	-	-	Low Expr
rs2736337	8	11341880	T	C	0.62	0.63	0.90	5E-12	BLK	-	-	-	-	-	rs998683	G/A	A	-23.3	#####	Low Expr
rs10985070	9	123636121	A	C	0.52	0.54	0.92	2E-09	TRAF1	rs10985070	C/A	C	-18.0	4E-72	-	-	-	-	-	Low Expr
rs2671692	10	50097819	A	G	0.57	0.53	1.08	9E-08	WDFY4	rs2671692	G/A	G	5.9	3E-09	-	-	-	-	-	Low Expr
rs968567	11	61595564	T	C	0.16	0.18	0.90	7E-07	FADS1	rs968567	T/C	T	16.6	8E-62	-	-	-	-	-	Low Expr
rs10774624	12	111833788	A	G	0.49	0.51	0.92	2E-07	SH2B3	-	-	-	-	-	rs4766578	T/A	T	9.2	6E-20	High Expr
rs4780401	16	11839326	T	G	0.57	0.56	1.07	6E-07	TXNDC11	-	-	-	-	-	rs8058003	T/C	C	6.0	2E-09	Low Expr
rs1877030	17	37740161	T	C	0.15	0.16	0.90	3E-08	IKZF3	rs1877030	C/T	T	-8.6	8E-18	-	-	-	-	-	High Expr
rs2469434	18	67544046	T	C	0.59	0.60	0.94	6E-06	CD226	-	-	-	-	-	rs763362	G/A	G	-12.2	4E-34	Low Expr
rs4239702	20	44749251	T	C	0.28	0.31	0.89	9E-15	CD40	rs4239702	T/C	T	-12.3	1E-34	-	-	-	-	-	High Expr
rs1893592	21	43855067	A	C	0.74	0.73	1.11	4E-12	UBASH3A	rs1893592	C/A	C	20.3	6E-92	-	-	-	-	-	Low Expr
rs909685	22	39747671	A	T	0.46	0.46	1.12	6E-14	SYNGR1	rs909685	T/A	A	25.3	#####	-	-	-	-	-	Low Expr

B. IBD-TRS (transcriptional risk score based on IBD GWAS, adapted from Supp. Table 4 in Marigorta et al. Nat Gen, 2017)

GWAS information Liu et al. Nat Gen, 2015)									Target Gene	eQTL activity at GWAS SNP (eQTLbloodbrowser)					eQTL activity at SNP in LD with GWAS SNP						
GWAS_SNP	chr	poshg19	A1	A2	A1_FreqCase	A1_FreqCtrl	OR	Pvalue	Target Gene	GWAS_SNP_eQTL_effect	GWAS_SNP_eQTL_effect_all	GWAS_SNP_eQTL_effect_effall	GWAS_SNP_eQTL_effect_effdir_zscore	GWAS_SNP_eQTL_effect_pval	GWAS_SNP_inLD_eQTL_effect	GWAS_SNP_inLD_eQTL_all	GWAS_SNP_inLD_eQTL_effall	GWAS_SNP_inLD_eQTL_effdir_zscore	GWAS_SNP_inLD_eQTL_pval	Inferred Direction of Risk
rs12103	1	1247494	A	G	-	-	1.09	1E-05	TNFRSF18	-	-	-	-	-	rs12142199	G/A	G	-9.4	3E-21	Low Expr
rs12103	1	1247494	A	G	-	-	1.09	1E-05	B3GALT6	-	-	-	-	-	rs12142199	G/A	G	-13.2	9E-40	Low Expr
rs12103	1	1247494	A	G	-	-	1.09	1E-05	SSU72	-	-	-	-	-	rs12142199	G/A	G	8.6	1E-17	High Expr
rs4845604	1	151801680	A	G	-	-	0.88	2E-05	THEM4	rs4845604	G/A	A	6.7	3E-11	-	-	-	-	-	Low Expr
rs4656958	1	160856964	A	G	-	-	0.93	4E-06	LY9	-	-	-	-	-	rs2184069	G/C	G	10.3	9E-25	Low Expr
rs4656958	1	160856964	A	G	-	-	0.93	4E-06	CD244	-	-	-	-	-	rs4656945	T/C	C	9.0	3E-19	Low Expr
rs1801274	1	161479745	G	A	-	-	0.88	3E-13	FCGR2B	rs1801274	A/G	G	-11.1	1E-28	-	-	-	-	-	High Expr
rs10185424	2	102662888	A	C	-	-	1.09	6E-09	IL1R2	rs10185424	T/G	T	-8.3	1E-16	-	-	-	-	-	Low Expr
rs2382817	2	219151218	A	C	-	-	1.08	4E-07	SLC11A1	rs2382817	C/A	A	14.7	8E-49	-	-	-	-	-	High Expr
rs9868809	3	48681053	A	G	-	-	1.15	6E-11	NCKIPSD	rs9868809	C/T	T	-7.1	1E-12	-	-	-	-	-	Low Expr
rs3197999	3	49721532	A	G	-	-	1.18	2E-21	USP4	-	-	-	-	-	rs1800668	G/A	A	12.8	1E-37	High Expr
rs2930047	5	10695526	G	A	-	-	1.08	1E-04	DAP	rs2930047	T/C	C	-16.5	5E-61	-	-	-	-	-	Low Expr
rs1363907	5	96252803	A	G	-	-	1.08	2E-07	LNPEP	rs1363907	G/A	A	15.5	3E-54	-	-	-	-	-	High Expr

rs11743851	5	130613600	G	A	-	-	1.11	4E-07	CDC42SE2	rs11743851	T/C	C	9.7	2E-22	-		-	-	-	-	High Expr
rs17622378	5	131778452	G	A	-	-	1.15	1E-15	SLC22A4	rs17622378	A/G	G	-23.9	#####	-		-	-	-	-	Low Expr
rs17622378	5	131778452	G	A	-	-	1.15	1E-15	SLC22A5	rs17622378	A/G	G	-27.0	#####	-		-	-	-	-	Low Expr
rs17622378	5	131778452	G	A	-	-	1.15	1E-15	IRF1	rs17622378	A/G	G	5.6	3E-08	-		-	-	-	-	High Expr
rs9313808	5	158820844	A	G	-	-	0.87	1E-12	RNF145	rs9313808	G/A	A	-6.7	2E-11	-		-	-	-	-	High Expr
rs4976646	5	176788570	G	A	-	-	1.08	4E-04	RGS14	rs4976646	T/C	C	5.2	2E-07	-		-	-	-	-	High Expr
rs17057051	8	27227554	G	A	-	-	0.94	9E-04	PTK2B	rs17057051	A/G	G	12.6	3E-36	-		-	-	-	-	Low Expr
rs4246905	9	117553249	A	G	-	-	0.88	1E-16	TNFSF8	rs4246905	C/T	T	12.6	2E-36	-		-	-	-	-	Low Expr
rs10781499	9	139266405	A	G	-	-	1.17	4E-25	CARD9	rs10781499	G/A	A	38.8	#####	-		-	-	-	-	High Expr
rs10781499	9	139266405	A	G	-	-	1.17	4E-25	SNAPC4	-	-	-	-	-	rs10781518	G/A	G	-4.2	3E-05	Low Expr
rs10781499	9	139266405	A	G	-	-	1.17	4E-25	SDCCAG3	rs10781499	G/A	A	-14.0	2E-44	-		-	-	-	-	Low Expr
rs10781499	9	139266405	A	G	-	-	1.17	4E-25	INPP5E	rs10781499	G/A	A	-34.4	#####	-		-	-	-	-	Low Expr
rs11230563	11	60776209	A	G	-	-	0.92	3E-06	SLC15A3	rs11230563	T/C	T	5.0	5E-07	-		-	-	-	-	Low Expr
rs559928	11	64150370	A	G	-	-	0.91	3E-05	RPS6KA4	rs559928	C/T	T	9.1	1E-19	-		-	-	-	-	Low Expr
rs8005161	14	88472595	A	G	-	-	1.15	4E-07	GALC	rs8005161	C/T	T	-8.9	6E-19	-		-	-	-	-	Low Expr
rs26528	16	28517709	G	A	-	-	1.1	2E-09	SBK1	-	-	-	-	-	rs4788084	T/C	T	-8.3	9E-17	Low Expr
rs26528	16	28517709	G	A	-	-	1.1	2E-09	CCDC101	-	-	-	-	-	rs4788084	T/C	T	-16.1	3E-58	Low Expr
rs727088	18	67530439	G	A	-	-	1.06	9E-03	CD226	rs727088	G/A	G	-14.4	1E-46	-		-	-	-	-	Low Expr
rs12720356	19	10469975	C	A	-	-	1.16	7E-11	ICAM4	rs12720356	A/C	C	7.4	1E-13	-		-	-	-	-	High Expr
rs12720356	19	10469975	C	A	-	-	1.16	7E-11	TYK2	rs12720356	A/C	C	4.7	2E-06	-		-	-	-	-	High Expr
rs11879191	19	10512911	A	G	-	-	0.89	5E-07	ICAM3	rs11879191	G/A	A	-6.7	2E-11	-		-	-	-	-	High Expr
rs913678	20	48955424	G	A	-	-	0.93	8E-06	CEBPB	rs913678	T/C	C	-8.6	9E-18	-		-	-	-	-	High Expr
rs6062504	20	62348907	A	G	-	-	0.9	3E-15	ZGPAT	-	-	-	-	-	rs6062509	T/G	G	-15.0	7E-51	High Expr
rs6062504	20	62348907	A	G	-	-	0.9	3E-15	LIME1	-	-	-	-	-	rs6011066	A/G	G	23.1	#####	Low Expr
rs2413583	22	39659773	A	G	-	-	0.84	2E-12	PDGFB	rs2413583	C/T	T	-4.8	2E-06	-		-	-	-	-	High Expr
rs12627970	22	39721745	G	A	-	-	1.12	4E-07	SYNGR1	rs12627970	A/G	G	21.5	#####	-		-	-	-	-	High Expr

Supplementary Table 2 – List of top 10 differentially expressed genes for various comparisons

Control v. Oligoarticular JIA					
Upregulated in Control			Upregulated in Oligoarticular JIA		
Gene	logFC	P-value	Gene	logFC	P-value
VPS26B	0.21	2.62E-04	CHRM3AS2	-1.00	1.05E-05
COPS7A	0.22	2.69E-04	IL6ST	-0.59	1.61E-04
HENMT1	0.27	3.54E-04	LINC01003	-0.55	3.07E-04
SIRT2	0.28	2.92E-05	SVIP	-0.51	1.33E-04
PIAS3	0.28	4.37E-05	PDK1	-0.46	3.05E-06
FAM50A	0.28	2.34E-04	ZNF518B	-0.31	1.58E-04
ITGAL	0.30	2.31E-04	CYLD	-0.29	3.14E-04
MRPL10	0.31	2.79E-04			
GALE	0.35	1.86E-04			
TBCB	0.36	1.29E-04			
Control v. Polyarticular JIA					
Upregulated in Control			Upregulated in Oligoarticular JIA		
Gene	logFC	P-value	Gene	logFC	P-value
BZRAP1	1.03	1.99E-08	SLC25A16	-0.37862	1.38E-05
ERBB2	1.01	1.56E-07	ELOVL7	-1.14234	1.56E-05
PLEKHF1	1.05	2.01E-07	PCYT1B	-0.97186	4.47E-05
NKG7	1.10	2.37E-07	PLA2G12A	-0.52328	6.70E-05
ARVCF	1.17	2.67E-07	PRRG4	-1.0844	6.76E-05
COL6A2	1.27	2.71E-07	SH3BGRL2	-0.94715	9.03E-05

PCDHGB1	1.07	3.68E-07	PRKAR2B	-0.97044	0.00011
NMUR1	1.20	3.89E-07	PDK1	-0.3763	0.00012
DLG5	1.04	6.10E-07	USP15MIR6125	-0.67584	0.00012
SCART1	1.36	6.79E-07	SCRN3	-0.38791	0.00013
Control v. Systemic JIA					
Upregulated in Control			Upregulated in Systemic JIA		
Gene	logFC	P-value	Gene	logFC	P-value
SCART1	2.00	6.16E-11	NFKBIZNXPE3	-0.83	2.22E-09
LTK	1.59	4.28E-10	NBN	-1.09	1.46E-08
DLG5	1.41	7.33E-10	MIR6502IRAK3	-1.95	4.76E-08
BZRAP1	1.24	8.33E-10	TNIP1	-0.91	2.94E-07
PCDHGB1	1.39	2.21E-09	GK	-1.59	3.12E-07
COLQ	1.16	5.25E-09	CD274	-2.37	6.51E-07
LGR6	1.88	5.27E-09	ST6GALNAC3	-1.62	6.55E-07
ADRB2	0.77	5.85E-09	ACSL4	-1.09	7.30E-07
NMUR1	1.50	1.34E-08	NAB1	-0.67	8.63E-07
NCR3	1.09	4.75E-08	SOD2	-1.32	8.89E-07
IBD v. Oligoarticular JIA					
Upregulated in IBD			Upregulated in Oligoarticular JIA		
Gene	logFC	P-value	Gene	logFC	P-value
GYG1	1.46	8.28E-17	QSOX2	-0.56	1.11E-17
PSENEN	0.61	5.56E-16	TTC39B	-0.68	2.92E-16
CST7	1.54	1.13E-15	NKD1	-1.09	4.25E-16

MMP8	3.94	1.19E-15	CTC1	-0.44	7.66E-16
TIMP1	0.87	1.82E-15	DNHD1	-0.73	1.91E-15
TXN	1.00	2.92E-15	CFAP44	-0.69	5.61E-15
HP	2.82	3.14E-15	ZNF550	-0.59	9.02E-15
S100A12	2.01	3.87E-15	ZBTB40	-0.48	1.01E-14
GLIPR2	0.71	4.05E-15	KLHL3	-0.76	1.30E-14
CLIC1	0.68	5.47E-15	CUBN	-0.86	1.55E-14
IBD v. Polyarticular JIA					
Upregulated in IBD			Upregulated in Polyarticular JIA		
Gene	logFC	P-value	Gene	logFC	P-value
CD177	4.56	1.73E-15	PRPF3	-0.32	3.14E-14
GYG1	1.29	2.55E-14	ZMAT1	-0.62	3.00E-13
MCEMP1	2.01	1.93E-13	NPIPB3	-0.66	5.23E-13
MMP8	3.40	2.39E-13	DDX26B	-0.42	5.61E-13
LRPAP1	0.57	5.00E-13	PROX2	-0.57	1.43E-12
SEPHS2	0.34	2.60E-12	CFAP44	-0.61	1.80E-12
PFKFB3	1.21	2.76E-12	NKTR	-0.54	2.00E-12
FCGR1A	1.61	4.06E-12	C3	-1.31	2.31E-12
CST7	1.28	5.48E-12	MASP2	-0.67	2.98E-12
ATP9A	1.31	1.53E-11	CTC1	-0.37	9.59E-12
IBD v. Systemic JIA					
Upregulated in IBD			Upregulated in Systemic JIA		
Gene	logFC	P-value	Gene	logFC	P-value

ALDH1A1	0.90	1.07E-09	ZC3H12A	-0.93	2.63E-18
C9orf47S1PR3	0.68	1.69E-07	TNFAIP3	-0.82	1.71E-14
CEBPA	0.60	4.19E-07	GHRL	-1.06	1.76E-10
CRTAP	0.40	5.51E-07	NFKB2	-0.56	3.67E-10
CPVL	0.65	5.83E-07	TTC21A	-0.98	1.30E-09
IL17RC	0.64	6.21E-07	PPM1N	-0.97	2.72E-09
PID1	0.82	6.34E-07	C3	-1.29	2.95E-09
MIR4709NPC2	0.46	1.68E-06	NFKBIZNXPE3	-0.49	6.00E-09
FOLR2	0.69	3.01E-06	USP18	-1.77	1.50E-08
ANAPC15	0.44	3.66E-06	TPK1	-0.44	1.80E-08

Supplementary Table 3 – Mean Blood Transcript Modules and BIT Axis Scores across Disease Subtypes

BTM	CTRL	IBD_CD	IBD_UC	JIA_Olig	JIA_Poly	JIA_Syst
integrin cell surface interactions (I)	0.992	-0.326	-0.492	0.479	-0.041	-0.130
integrin cell surface interactions (II)	0.288	0.389	-0.015	-0.072	-0.175	-0.459
extracellular matrix (I)	1.048	-0.341	-0.714	0.667	-0.066	-0.243
extracellular matrix (II)	-0.545	0.384	0.750	-0.583	-0.158	0.236
extracellular matrix (III)	0.723	-0.516	-0.699	0.692	0.075	-0.069
regulation of signal transduction	0.650	-0.359	-0.650	0.557	0.078	-0.189
cell cycle and transcription	0.205	0.189	-0.100	-0.086	0.129	-0.531
cell cycle (I)	-0.389	-0.012	0.368	-0.139	-0.210	0.604
PLK1 signaling events	-0.465	0.046	0.401	-0.191	-0.233	0.621
myeloid cell enriched receptors and transporters	-0.434	0.461	0.596	-0.583	-0.093	-0.011
mitotic cell cycle - DNA replication	-0.211	-0.087	0.098	-0.036	-0.162	0.574
mitotic cell cycle in stimulated CD4 T cells	-0.479	0.011	0.351	-0.151	-0.191	0.583
cell division in stimulated CD4 T cells	-0.587	0.022	0.353	-0.164	-0.185	0.612
mitotic cell cycle	-0.379	0.013	0.419	-0.164	-0.214	0.569
cell division - E2F transcription network	0.304	-0.015	0.095	0.144	-0.337	0.285
cell cycle (II)	-0.434	0.076	0.489	-0.210	-0.227	0.521
C-MYC transcriptional network	-0.433	0.043	0.413	-0.117	-0.263	0.556
cell junction (GO)	-0.528	0.529	0.893	-0.623	-0.165	-0.063
Rho GTPase cycle	-0.601	0.053	0.479	-0.171	-0.239	0.603
enriched in monocytes (I)	-0.120	0.531	0.736	-0.591	-0.126	-0.274
regulation of antigen presentation and immune response	-0.672	0.359	0.679	-0.568	-0.080	0.203
T cell activation and signaling	0.610	-0.214	-0.660	0.453	0.019	-0.193

mitotic cell division	-0.487	-0.029	0.365	-0.115	-0.207	0.638
enriched in T cells (I)	0.714	-0.230	-0.813	0.402	0.065	-0.137
T cell activation (I)	0.470	-0.308	-0.756	0.510	0.069	-0.078
enriched in NK cells (I)	1.345	0.028	-0.922	0.011	0.007	-0.162
T cell activation (II)	0.893	-0.166	-0.849	0.323	0.051	-0.176
T cell activation (III)	0.210	-0.385	-0.701	0.508	0.143	0.016
E2F transcription factor network	-0.517	0.049	0.547	-0.255	-0.183	0.562
B cell development	-0.586	0.405	0.556	-0.448	-0.130	0.052
E2F1 targets (Q3)	-0.342	-0.076	0.101	-0.026	-0.131	0.529
E2F1 targets (Q4)	-0.330	-0.086	0.070	-0.081	-0.105	0.592
enriched in monocytes (II)	-0.606	0.456	0.705	-0.580	-0.117	0.056
blood coagulation	-0.553	0.431	0.673	-0.600	-0.112	0.120
formyl peptide receptor mediated neutrophil response	-0.553	0.413	0.457	-0.496	-0.118	0.122
CD28 costimulation	-0.496	0.267	0.561	-0.510	-0.013	0.165
innate activation by cytosolic DNA sensing	-0.451	0.309	0.028	-0.450	0.002	0.210
T cell differentiation	0.084	-0.380	-0.653	0.603	0.119	-0.066
Ran mediated mitosis	-0.203	0.133	0.405	-0.174	-0.311	0.458
TLR and inflammatory signaling	-0.658	0.371	0.623	-0.515	-0.104	0.168
Hox cluster I	-0.454	0.099	0.637	-0.485	0.171	0.065
T cell differentiation via ITK and PKC	0.272	-0.312	-0.696	0.461	0.097	-0.002
T cell differentiation (Th2)	0.355	-0.243	-0.600	0.449	0.057	-0.130
AP-1 transcription factor network	-0.480	0.359	0.669	-0.230	-0.287	0.017
cell adhesion (lymphocyte homing)	0.633	-0.342	-0.747	0.468	0.126	-0.126
mismatch repair (I)	-0.346	-0.192	-0.006	0.119	-0.001	0.354
mismatch repair (II)	-0.200	-0.220	-0.216	0.219	0.010	0.287
RA, WNT, CSF receptors network (monocyte)	-0.329	0.342	0.511	-0.371	-0.139	0.002
cell activation (IL15, IL23, TNF)	-0.465	0.294	0.167	-0.367	-0.137	0.315
TLR8-BAFF network	-0.742	0.397	0.709	-0.565	-0.098	0.169

TBA	-0.663	-0.259	-0.104	0.111	0.238	0.225
chemokine cluster (I)	1.206	-0.066	-0.692	-0.036	0.008	0.036
chemokine cluster (II)	0.989	0.247	-0.655	-0.426	0.080	-0.082
antigen presentation (lipids and proteins)	0.863	-0.220	-0.788	0.386	-0.024	-0.037
proinflammatory cytokines and chemokines	0.258	0.074	-0.028	-0.472	0.114	0.251
cell movement, Adhesion & Platelet activation	-0.888	0.297	0.406	-0.344	-0.013	0.088
cell cycle and growth arrest	-0.348	0.374	0.398	-0.397	-0.206	0.167
platelet activation (I)	0.317	0.456	0.476	-0.275	-0.378	-0.161
platelet activation (II)	0.256	0.413	0.409	-0.259	-0.313	-0.162
CORO1A-DEF6 network (I)	0.555	0.216	0.031	0.084	-0.311	-0.213
KLF12 targets network	0.268	0.353	0.368	-0.195	-0.328	-0.095
CORO1A-DEF6 network (II)	0.472	0.251	0.075	0.019	-0.294	-0.206
cytoskeletal remodeling	0.120	0.416	0.324	-0.297	-0.218	-0.195
inflammatory response	-0.174	0.486	0.703	-0.591	-0.193	-0.018
cytoskeletal remodeling (enriched for SRF targets)	-0.883	0.365	0.464	-0.422	-0.076	0.157
signaling in T cells (I)	1.166	0.084	-0.798	-0.127	0.050	-0.143
signaling in T cells (II)	1.318	0.096	-0.805	-0.024	-0.061	-0.168
T cell surface, activation	0.256	-0.220	-0.635	0.493	0.027	-0.127
immune activation - generic cluster	-0.678	0.388	0.710	-0.543	-0.113	0.158
enriched in neutrophils (I)	-0.653	0.331	0.577	-0.414	-0.119	0.148
endoplasmic reticulum	0.891	-0.338	-0.647	0.522	0.050	-0.223
cell division	-0.025	-0.140	-0.121	0.229	-0.095	0.191
chemokines and receptors	1.158	-0.216	-0.744	0.214	0.149	-0.255
integrin mediated leukocyte migration	-0.255	0.467	0.803	-0.556	-0.247	0.056
complement and other receptors in DCs	-0.383	0.526	0.697	-0.702	-0.173	0.116
platelet activation (III)	-0.519	0.364	0.613	-0.486	-0.150	0.175
myeloid, dendritic cell activation via NFkB (I)	-0.252	0.317	0.304	-0.392	-0.209	0.291
myeloid, dendritic cell activation via NFkB (II)	-0.325	0.367	0.438	-0.405	-0.230	0.210

T cell signaling and costimulation	-0.641	0.487	0.782	-0.635	-0.172	0.154
leukocyte activation and migration	-0.191	0.319	0.649	-0.515	-0.143	0.147
cell division stimulated CD4+ T cells	-0.526	0.082	0.512	-0.195	-0.256	0.566
enriched in B cells (I)	0.093	-0.315	-0.436	0.417	0.012	0.174
enriched in B cells (II)	0.103	-0.317	-0.338	0.420	-0.018	0.176
enriched in B cells (III)	-0.293	0.356	0.438	-0.472	-0.021	-0.079
enriched in B cells (IV)	0.059	-0.359	-0.384	0.481	0.009	0.161
enriched in B cells (V)	0.492	-0.456	-0.766	0.491	0.175	0.044
transcription regulation in cell development	-0.622	0.418	0.694	-0.561	-0.153	0.183
CD1 and other DC receptors	-0.866	0.154	0.635	-0.270	-0.091	0.293
cell adhesion	-0.708	0.287	0.292	-0.309	-0.011	0.037
T cell activation (IV)	0.769	-0.241	-0.606	0.375	0.064	-0.198
inflammasome receptors and signaling	-0.618	0.387	0.383	-0.538	-0.085	0.237
BCR signaling	0.256	-0.333	-0.534	0.393	0.206	-0.139
suppression of MAPK signaling	-0.495	0.377	0.526	-0.464	-0.147	0.142
immuregulation - monocytes, T and B cells	0.146	-0.286	-0.556	0.441	0.004	0.132
B cell development/activation	0.438	-0.385	-0.502	0.507	0.055	-0.006
CCR1, 7 and cell signaling	-0.618	0.366	0.656	-0.539	-0.081	0.137
lymphocyte generic cluster	0.406	-0.326	-0.704	0.482	0.090	-0.041
enriched in NK cells (II)	1.430	0.033	-0.877	0.007	-0.040	-0.136
enriched in NK cells (KIR cluster)	0.905	0.042	-0.491	-0.163	-0.076	0.185
enriched in NK cells (receptor activation)	1.284	-0.135	-0.973	0.226	-0.004	-0.084
T & B cell development, activation	0.249	-0.287	-0.615	0.544	0.038	-0.097
enriched for unknown TF motif CTCNANGTGNY	0.871	0.014	-0.276	0.312	-0.114	-0.491
regulation of localization (GO)	0.397	0.403	0.251	-0.195	-0.324	-0.195
enriched in activated dendritic cells/monocytes	-0.668	0.327	0.488	-0.448	-0.109	0.237
IL2, IL7, TCR network	0.260	-0.266	-0.582	0.524	0.027	-0.110
activated dendritic cells	0.584	-0.231	0.282	0.389	-0.187	-0.132

RIG-1 like receptor signaling	-0.189	0.098	-0.565	-0.218	0.230	0.051
enriched in B cells (VI)	0.034	-0.332	-0.476	0.407	0.068	0.160
enriched in antigen presentation (I)	-0.972	0.232	0.648	-0.547	0.052	0.315
enriched in monocytes (III)	-0.737	0.371	0.561	-0.524	-0.051	0.142
transcriptional targets of glucocorticoid receptor	0.165	-0.368	-0.625	0.416	0.188	0.015
antiviral IFN signature	-0.383	0.134	-0.525	-0.321	0.266	0.119
DNA repair	-0.346	-0.237	-0.246	0.021	0.118	0.488
collagen, TGFB family et al	0.508	-0.464	-0.740	0.488	0.224	-0.047
myeloid cell cytokines, metallopeptidases and laminins	0.616	-0.404	-0.746	0.592	0.155	-0.234
enriched in myeloid cells and monocytes	-0.867	0.378	0.647	-0.539	-0.030	0.120
signal transduction, plasma membrane	0.895	-0.229	-0.381	0.340	-0.017	-0.180
enriched in naive and memory B cells	-0.009	-0.335	-0.447	0.380	0.097	0.154
integrins and cell adhesion	-0.490	0.348	0.635	-0.502	-0.125	0.160
platelet activation and degranulation	-0.551	0.436	0.471	-0.561	-0.046	0.027
chemokines and inflammatory molecules in myeloid cells	-0.622	0.262	0.284	-0.424	-0.105	0.411
proinflammatory dendritic cell, myeloid cell response	0.765	-0.241	-0.395	0.441	0.002	-0.285
transmembrane transport (I)	-0.585	0.394	0.679	-0.595	-0.113	0.201
leukocyte migration	0.885	-0.331	-0.809	0.559	0.082	-0.270
putative targets of PAX3	0.818	-0.055	-0.548	0.117	0.044	-0.202
adhesion and migration, chemotaxis	1.174	0.306	-0.334	-0.465	0.015	-0.255
lipid metabolism, endoplasmic reticulum	-0.941	0.326	0.710	-0.502	0.009	0.094
growth factor induced, enriched in nuclear receptor subfamily 4	0.632	-0.260	-0.356	0.483	0.050	-0.365
enriched in antigen presentation (II)	0.644	-0.400	-0.603	0.602	0.073	-0.187
enriched in antigen presentation (III)	-0.626	0.419	0.512	-0.598	-0.098	0.226
enriched for SMAD2/3 signaling	0.553	-0.219	-0.208	0.251	-0.054	0.053
MAPK, RAS signaling	0.044	0.392	0.361	-0.272	-0.252	-0.110
phosphatidylinositol signaling system	-1.061	0.002	0.223	-0.159	0.136	0.288
cell cycle (III)	-0.293	-0.019	0.319	-0.125	-0.201	0.565

nuclear pore complex	-0.050	-0.338	-0.433	0.365	0.156	0.080
nuclear pore complex (mitosis)	-0.244	-0.290	-0.405	0.362	0.109	0.140
receptors, cell migration	0.598	-0.343	-0.765	0.512	0.112	-0.142
axon guidance	0.895	0.366	0.152	-0.372	-0.113	-0.405
viral sensing & immunity; IRF2 targets network (I)	-0.371	0.541	0.305	-0.610	-0.098	-0.003
viral sensing & immunity; IRF2 targets network (II)	-0.403	0.372	-0.129	-0.461	0.069	0.030
complement activation (I)	-0.356	0.590	0.616	-0.720	-0.158	0.015
complement activation (II)	-0.790	0.372	0.478	-0.514	0.003	0.087
golgi membrane (I)	-0.998	0.329	0.740	-0.503	-0.054	0.217
glycerophospholipid metabolism	0.935	-0.399	-0.771	0.537	0.060	-0.094
cytokines - recepters cluster	-0.335	0.476	0.744	-0.573	-0.179	-0.007
cell adhesion (GO)	0.430	-0.373	-0.612	0.536	0.122	-0.142
enriched in monocytes (IV)	-0.774	0.369	0.672	-0.540	-0.070	0.167
enriched in monocytes (surface)	0.647	-0.409	-0.792	0.564	0.152	-0.166
enriched in activated dendritic cells (I)	-0.469	0.378	0.622	-0.691	0.069	0.013
enriched for cell migration	0.818	0.020	-0.842	-0.013	0.155	-0.221
enriched in B cell differentiation	0.584	-0.256	-0.501	0.547	-0.113	-0.068
enriched in membrane proteins	-0.586	0.363	0.782	-0.525	-0.222	0.315
double positive thymocytes	0.546	-0.321	-0.731	0.556	0.082	-0.190
type I interferon response	-0.515	0.158	-0.447	-0.329	0.222	0.179
inositol phosphate metabolism	-1.017	-0.078	0.163	-0.026	0.072	0.386
enriched in G-protein coupled receptors	1.252	0.109	-0.669	-0.205	0.070	-0.213
recruitment of neutrophils	-0.623	0.354	0.510	-0.416	-0.119	0.125
cell adhesion, membrane	1.372	-0.237	-0.788	0.381	0.087	-0.409
Membrane, ER proteins	-0.287	0.167	0.563	-0.416	-0.018	0.152
enriched for ubiquitination	-0.717	-0.139	-0.009	0.132	0.104	0.179
lysosomal/endosomal proteins	-0.267	0.535	0.834	-0.640	-0.182	-0.094
extracellular matrix, complement	-0.412	0.426	0.695	-0.515	-0.207	0.115

nuclear pore, transport; mRNA splicing, processing	-0.112	-0.329	-0.490	0.396	0.179	0.025
cell cycle, ATP binding	-1.163	0.062	0.530	-0.264	0.105	0.269
cytoskeleton/actin (SRF transcription targets)	-0.906	0.283	0.389	-0.431	0.062	0.122
intracellular transport	-0.989	-0.043	0.167	-0.036	0.106	0.252
innate antiviral response	-0.490	0.128	-0.363	-0.395	0.270	0.196
amino acid metabolism and transport	-0.317	0.413	0.572	-0.441	-0.103	-0.141
G protein coupled receptors cluster	0.724	-0.302	-0.685	0.420	0.158	-0.263
plasma cells & B cells, immunoglobulins	0.171	-0.330	-0.361	0.446	-0.026	0.164
plasma cells, immunoglobulins	0.324	-0.052	0.257	0.268	-0.288	-0.008
enriched in NK cells (III)	1.101	-0.137	-0.983	0.240	0.011	-0.052
G protein mediated calcium signaling	0.840	-0.379	-0.616	0.457	0.008	0.043
leukocyte differentiation	-0.147	0.246	0.298	-0.230	-0.195	0.123
plasma membrane, cell junction	1.147	-0.281	-1.047	0.494	0.152	-0.394
cell junction	1.104	-0.021	-0.488	0.343	-0.127	-0.429
enriched in neutrophils (II)	-0.480	0.340	0.475	-0.429	-0.070	0.038
enriched in activated dendritic cells (II)	-0.789	0.242	-0.059	-0.486	0.209	0.185
enriched in cell cycle	-0.967	-0.093	0.126	-0.154	0.214	0.338
enriched in dendritic cells	0.801	-0.423	-0.970	0.616	0.138	-0.160
mitosis (TF motif CCAATNNSNNNGCG)	-0.430	-0.308	-0.200	0.212	0.242	0.121
heme biosynthesis (I)	-0.367	0.197	0.340	-0.270	-0.164	0.295
enriched for TF motif TTCNRGNNNNTTC	-0.470	0.542	0.974	-0.415	-0.202	-0.394
erythrocyte differentiation	-0.378	0.244	0.435	-0.293	-0.183	0.223
cell development	-0.115	0.342	0.369	-0.641	0.029	0.073
enriched for promoter motif NATCACGTGAY (putative SREBF1 targets)	0.176	-0.199	-0.414	0.389	-0.103	0.152
enriched for TF motif PAX3	-0.520	-0.320	-0.291	0.342	0.257	0.003
nucleotide metabolism	0.116	-0.236	-0.289	0.365	-0.040	0.105
enriched in DNA interacting proteins	0.155	-0.206	-0.362	0.227	0.015	0.168
extracellular region cluster (GO)	-0.444	0.107	0.566	-0.632	0.191	0.262

transmembrane transport (II)	-0.479	-0.294	-0.274	0.242	0.221	0.144
muscle contraction, SRF targets	-0.517	0.058	-0.077	-0.309	0.185	0.238
platelet activation - actin binding	-0.885	0.331	0.367	-0.395	0.054	-0.012
platelet activation & blood coagulation	-0.808	0.305	0.382	-0.339	0.031	-0.039
chaperonin mediated protein folding (I)	0.152	-0.018	0.092	0.139	-0.082	-0.122
chaperonin mediated protein folding (II)	-0.013	-0.003	0.158	0.141	-0.102	-0.086
Wnt signaling pathway	0.016	0.254	-0.279	0.025	-0.036	-0.353
lysosome	-0.324	0.549	0.842	-0.702	-0.148	-0.073
extracellular matrix, collagen	-0.257	0.448	0.685	-0.619	-0.133	0.031
purine nucleotide biosynthesis	0.353	-0.066	-0.103	0.232	-0.130	-0.058
regulation of transcription, transcription factors	-0.858	-0.322	-0.212	0.055	0.348	0.359
small GTPase mediated signal transduction	-0.138	0.363	0.453	-0.385	-0.192	0.033
respiratory electron transport chain (mitochondrion)	0.289	0.463	0.640	-0.242	-0.400	-0.254
heme biosynthesis (II)	-0.245	0.185	0.514	-0.276	-0.157	0.179
enriched in T cells (II)	0.351	-0.390	-0.902	0.537	0.173	-0.032
proteasome	-0.324	0.368	0.613	-0.385	-0.188	0.010
translation initiation	-0.135	0.226	0.636	-0.259	-0.162	-0.028
olfactory receptors	0.503	-0.418	-0.646	0.201	0.374	-0.040
cell cycle, mitotic phase	-0.109	-0.376	-0.563	0.405	0.246	0.014
enriched for TF motif TNCATNTCCYR	-0.035	0.018	-0.257	0.044	-0.043	0.122
transcription elongation, RNA polymerase II	0.512	0.441	0.493	-0.143	-0.449	-0.280
mitochondrial cluster	0.228	-0.145	-0.050	0.253	-0.033	-0.088
golgi membrane (II)	-1.115	0.045	0.334	-0.200	0.104	0.290
chromosome Y linked	-0.076	-0.002	0.388	-0.085	-0.204	0.356
translation initiation factor 3 complex	0.390	0.136	0.064	0.163	-0.228	-0.274
spliceosome	-0.101	0.148	0.252	-0.167	-0.061	-0.020
T cell surface signature	0.340	-0.267	-0.629	0.486	0.048	-0.099
NK cell surface signature	1.360	0.045	-0.750	-0.094	-0.012	-0.096

B cell surface signature	0.236	-0.417	-0.535	0.477	0.083	0.147
Plasma cell surface signature	0.288	0.261	0.747	-0.092	-0.359	-0.200
Monocyte surface signature	-0.658	0.385	0.692	-0.539	-0.106	0.144
DC surface signature	0.931	-0.305	-0.688	0.463	0.108	-0.306
CD4 T cell surface signature Th1-stimulated	0.990	-0.095	-0.556	0.237	-0.007	-0.265
CD4 T cell surface signature Th2-stimulated	0.030	-0.333	-0.478	0.458	0.110	0.008
Naive B cell surface signature	-0.066	-0.308	-0.488	0.336	0.165	0.078
Memory B cell surface signature	-0.442	0.114	-0.334	-0.372	0.289	0.123
Resting dendritic cell surface signature	1.053	-0.235	-0.701	0.467	0.033	-0.356
Activated (LPS) dendritic cell surface signature	-0.823	0.302	0.322	-0.451	0.007	0.217
Platelets	-0.769	0.305	0.392	-0.335	0.044	-0.091
Interferon	-0.466	0.277	-0.240	-0.420	0.122	0.141
Cell Cycle	0.151	-0.312	-0.399	0.351	0.122	0.014
Erythrocytes	-0.367	0.250	0.402	-0.316	-0.165	0.226
Inflammation	-0.704	0.400	0.684	-0.522	-0.134	0.164
Cytotoxic/NK Cell	1.352	0.054	-0.904	0.012	-0.015	-0.188
T cell	0.063	-0.441	-0.649	0.649	0.153	-0.068
Protein Synthesis	0.323	0.079	-0.004	0.257	-0.170	-0.350
B cell	0.181	-0.314	-0.422	0.424	-0.001	0.142
Monocytes	0.141	0.362	0.427	-0.393	-0.016	-0.387
T cells	0.512	-0.339	-0.779	0.547	0.085	-0.107
Mitochondrial Stress / Proteasome	-0.163	-0.010	0.126	-0.029	0.048	-0.025
Mitochondrial Respiration	-0.845	-0.322	-0.350	0.072	0.374	0.347
Neutrophils	-0.095	0.305	0.490	-0.378	-0.260	0.234
Apoptosis / Survival	-0.267	0.503	0.831	-0.477	-0.298	-0.051
Mitochondrial Stress	0.528	-0.236	-0.334	0.515	-0.077	-0.189
Cell Death	-0.661	0.414	0.699	-0.554	-0.151	0.190
Cytotoxic/NK	1.231	0.017	-0.826	-0.072	0.011	-0.026

Immune Responses	-0.601	0.313	0.800	-0.644	0.012	0.146
Axis T (1)	0.655	0.010	-0.325	0.425	-0.257	-0.269
Axis R (2)	-0.447	0.207	0.311	-0.295	-0.130	0.291
Axis B (3)	0.265	-0.296	-0.364	0.398	-0.055	0.186
Axis G (4)	-0.991	0.129	0.364	-0.380	0.121	0.295
Axis N (5)	-0.462	0.393	0.609	-0.437	-0.182	0.083
Axis I (7)	-0.542	0.108	-0.616	-0.287	0.323	0.124
Axis C	-0.269	0.144	0.467	-0.150	-0.373	0.516

Supplementary Table 4 – List of disease-by-eQTL interactions

Gene	rsID	IBD β	IBD p-val	JIA β	JIA p-val	Interact p
SKI	rs3001167	-1.7	6.76E-06	0.163519	0.241677	1.64E-06
FASTKD2	rs74345488	0.407	0.207	-1.3235	2.17E-05	7.08E-05
FAM20C	rs80086012	-0.507	0.105	0.747802	2.75E-05	0.000165
CAMKK1	rs903506	-0.961	2.00E-05	0.01009	0.939037	0.000175
RAC2	rs6000618	0.457	0.0349	-0.63489	7.42E-05	0.000232
SIRT3	rs2043055	-0.694	9.68E-05	0.128827	0.333153	0.000268
SDHC	rs16832809	-2.15	3.87E-05	-0.1589	0.501788	0.000414
NDUFA3	rs7254645	-0.192	0.215	0.562053	8.35E-05	0.000602
USP14	rs8085585	-0.562	7.01E-05	-0.29858	0.151345	0.000649
ACP1	rs34894023	0.449	0.227	-1.0482	6.42E-05	0.000688
DCTN5	rs703767	0.227	0.195	1.020839	4.18E-09	0.00074
ATAD3B	rs4077630	0.0314	0.844	-0.57241	2.04E-05	0.000882
PRDX6	rs4279882	1.84	3.83E-05	0.363601	0.054769	0.000985
FN3K	rs12940475	-0.0784	0.687	0.734389	1.44E-05	0.001501
GRAP2	rs137981	-1.14	4.46E-05	-0.09066	0.666071	0.002091
SNRNP25	rs8061370	0.605	2.99E-05	-0.06293	0.696224	0.002163
BCKDK	rs2855475	-0.153	0.38	0.561202	9.02E-05	0.002216
THBS3	rs4971079	0.0511	0.768	-0.54813	5.18E-05	0.003202
KIAA1683	rs10854163	0.0462	0.805	-0.65634	6.97E-08	0.004201
PPOX	rs12745476	0.206	0.35	-0.4601	5.42E-05	0.005352
VPS28	rs1871538	-0.136	0.628	-1.19038	6.37E-05	0.008327
ORMDL3	rs1565923	1.11	8.79E-07	0.468678	0.000618	0.008641
CEACAM21	rs4802122	-1.33	2.06E-08	-0.64651	2.21E-05	0.009499
ARPC2	rs13429408	-0.816	6.55E-05	0.175865	0.217886	0.009998
ATAD3A	rs973608	-0.726	1.14E-06	0.212146	0.288602	0.01189
TBKBP1	rs4289035	1.06	4.90E-06	-0.39551	0.001875	0.013043
SMG7	rs7550777	0.748	3.21E-05	0.231347	0.069313	0.016011
MGAT3	rs5757680	0.397	0.0206	0.988105	1.40E-08	0.016343
GEN1	rs10497450	1.58	3.92E-05	0.270692	0.253225	0.016633
LOC100506124	rs717027	0.69	5.48E-05	0.230354	0.156074	0.018442

LOC728613	rs4958354	-0.774	4.79E-05	-0.12044	0.437221	0.01976
ATP11B	rs2314737	0.0528	0.786	0.654773	8.57E-06	0.026845
PNRC2	rs1938341	0.0319	0.861	-0.56093	8.06E-05	0.027235
TMEM180	rs2274351	-0.0751	0.687	-0.54095	3.44E-05	0.027943
RPL12	rs2254437	-0.069	0.751	-0.53718	3.12E-05	0.028571
DFNB31	rs2296262	0.319	0.0537	0.74847	1.56E-08	0.030896
DNMBP	rs7893702	0.0467	0.895	-0.71325	2.40E-05	0.039025
CPTP	rs11809901	-1.08	9.81E-05	-0.11297	0.696146	0.040807
DCUN1D1	rs2314737	0.221	0.195	0.630321	2.14E-07	0.048337
FAM195A	rs16954348	0.357	0.51	-1.05988	1.14E-06	0.050343
MILR1	rs17401012	1.81	3.40E-06	0.95454	0.000311	0.051946
FAM118A	rs136611	1.03	4.21E-05	1.545911	1.31E-20	0.054032
METTL18	rs12130372	-0.565	0.063	-1.38499	1.93E-05	0.054107
PPFIA4	rs6683283	0.165	0.394	0.803547	1.69E-05	0.05706
JMJD8	rs16954348	0.624	0.327	-0.83088	6.34E-05	0.06035
LOC100996324	rs12959287	0.16	0.41	0.595204	2.61E-05	0.060527
LGALS9	rs11080242	-0.927	1.01E-05	-0.48729	0.000518	0.06209
LGALS9	rs1984547	-0.884	2.36E-05	-0.55024	4.14E-05	0.157601
ADAM1A	rs16941831	1.23	1.92E-05	0.561453	0.014438	0.062618
GPN3	rs4766500	-0.806	5.22E-05	-0.51343	0.000135	0.062985
HEATR3	rs7186889	-0.51	0.00239	-0.89151	7.11E-13	0.072233
RNASET2	rs398278	-0.443	0.0108	-1.06841	3.30E-15	0.073802
SLC22A5	rs11950562	-0.534	8.00E-04	-0.86308	6.07E-14	0.074362
SERAC1	rs11756587	0.239	0.423	0.9104	1.95E-05	0.091806
C16orf13	rs16954348	0.312	0.591	-0.92939	9.97E-06	0.094278
AHI1	rs12206850	-0.567	0.00285	-0.9239	2.93E-11	0.100782
RBCK1	rs6081158	0.76	2.44E-05	-0.27994	0.076173	0.106272
CD226	rs12969613	0.633	2.16E-07	0.179081	0.145273	0.109736
LGALS9C	rs11869582	-0.331	0.189	-0.66791	3.05E-05	0.109737
PIGQ	rs16954348	0.253	0.639	-0.91244	5.22E-05	0.111246
C1QTNF6	rs73161818	0.901	1.97E-06	0.499751	0.002773	0.118617
C1QTNF6	rs5756558	0.413	0.0106	0.519604	7.28E-05	0.587274
CST7	rs12185807	0.0599	0.81	0.433896	7.67E-05	0.123766

B3GNT5	rs2314737	0.255	0.163	0.80748	3.54E-06	0.127941
FGFR1OP	rs56019630	1.33	6.73E-06	0.856966	1.67E-06	0.128701
WDR27	rs4498361	0.165	0.34	0.586065	3.36E-05	0.129419
EXTL2	rs17123491	-0.391	0.119	-0.82527	8.52E-05	0.142375
WDR24	rs16954348	0.236	0.702	-0.88567	4.66E-05	0.159592
VNN1	rs3798792	0.77	1.30E-05	0.502361	9.31E-06	0.167815
IL32	rs17790434	-0.0846	0.778	-0.5365	5.70E-05	0.174578
KLC1	rs729438	0.485	0.014	0.699987	1.37E-06	0.183549
LOC645513	rs13113112	-1.13	1.04E-07	-0.90461	2.59E-11	0.19021
LOC645513	rs6849889	-1.09	3.33E-07	-0.90465	1.03E-12	0.314709
MORN3	rs2720022	-0.279	0.157	0.478632	3.63E-05	0.19601
NTSR1	rs2427381	-0.786	0.00534	-0.48058	5.17E-05	0.207063
C4orf33	rs10518544	0.425	0.242	1.006358	8.73E-05	0.214409
SLC11A1	rs78846874	-0.35	0.36	-0.82787	3.85E-06	0.218666
CYP27A1	rs1516086	0.413	0.0236	-0.79083	7.84E-10	0.219749
NOD2	rs1981760	1.28	2.74E-08	1.045318	2.27E-16	0.233908
EIF2D	rs61048992	-0.505	0.146	-0.9922	1.72E-05	0.234918
RAB40C	rs1077352	-0.62	3.49E-05	-0.14382	0.343496	0.261467
ANAPC13	rs4368544	-0.345	0.0451	-0.65629	8.42E-06	0.262567
LINC01001	rs1440284	-0.284	0.228	-0.55358	9.25E-05	0.275624
ARPC5	rs10797891	-0.281	0.123	0.484582	9.73E-06	0.280105
UBE2E3	rs10195880	0.267	0.161	0.689888	8.70E-05	0.28372
NBR2	rs4534897	-0.817	0.000121	-0.99417	2.31E-15	0.300471
NADK	rs1475766	-0.664	5.90E-06	-0.21048	0.125321	0.302007
TMTC4	rs9585476	0.67	9.12E-05	0.864413	1.35E-10	0.31538
GPR35	rs2975788	0.373	0.0424	0.553759	3.18E-05	0.32937
DSE	rs9400918	-0.423	0.0282	-0.59882	2.79E-05	0.330555
GPD2	rs297581	0.593	8.98E-05	0.586601	0.000209	0.3329
C5	rs10818500	-0.39	0.0168	0.737297	3.08E-07	0.339861
C1orf85	rs2274226	1.15	1.89E-09	-1.01613	2.60E-21	0.344095
MIB2	rs12075213	-0.26	0.526	-0.83477	5.90E-05	0.358394
ZNF880	rs2059818	0.993	2.03E-08	0.960926	1.03E-10	0.371433
FCGR2B	rs10494360	0.626	0.0336	1.007047	1.80E-05	0.382129

MTRF1L	rs6557250	-0.708	5.19E-05	-0.91534	6.32E-06	0.40392
PNKD	rs13430006	0.344	0.137	0.565642	6.76E-07	0.40517
RPL3	rs137627	-0.272	0.1	0.504916	7.31E-05	0.409477
TDRD9	rs1187448	0.424	0.112	0.650718	1.17E-05	0.41899
NLRP2	rs12975582	0.562	0.0129	0.796243	1.19E-06	0.432793
SPPL3	rs12427353	0.861	1.62E-05	-0.20224	0.313791	0.44264
RAB11FIP3	rs16954348	-0.257	0.702	-0.84156	2.03E-05	0.446235
CFAP126	rs35006684	-0.412	0.2	-0.6577	2.11E-05	0.454833
PAM	rs2431321	1.04	3.77E-09	1.151941	2.10E-23	0.476862
LINC01061	rs13113112	0.98	3.46E-07	1.08365	4.07E-14	0.493742
LINC01061	rs7661020	1.05	1.48E-08	1.030043	2.36E-13	0.937769
CCDC127	rs422115	-0.873	5.14E-06	0.097156	0.583444	0.499223
ST7L	rs2999155	0.695	1.80E-05	0.833434	1.40E-07	0.508311
THBD	rs844885	-0.823	0.0258	-1.10479	3.29E-06	0.513719
KSR1	rs2945378	-0.476	0.00619	-0.59863	4.37E-07	0.522783
ST7L	rs6682737	0.695	1.80E-05	0.831406	1.07E-07	0.530326
YBEY	rs2839235	0.765	0.000646	0.697255	9.23E-08	0.540517
RMI2	rs11644184	0.581	0.000756	-0.69866	2.99E-07	0.540646
LOC613037	rs13331817	0.409	0.00789	0.62636	2.26E-05	0.547664
PLTP	rs7275164	-0.555	0.000206	-0.71355	6.95E-07	0.583535
CDK2AP1	rs11057223	1.29	8.09E-05	1.080669	1.73E-06	0.588468
CDK2AP1	rs12317452	1.29	8.94E-05	1.102224	6.43E-07	0.637829
PROCA1	rs8074623	-0.797	0.00504	-1.01211	3.52E-05	0.606873
PPIL3	rs12620435	-0.592	0.0171	-0.73336	4.99E-05	0.612963
GSN	rs7028970	0.437	0.0169	0.539064	8.45E-05	0.613692
ZNF467	rs855676	-0.532	0.027	-0.61017	4.81E-05	0.61607
GBAP1	rs914615	0.604	0.000316	0.803634	7.82E-10	0.617967
GBAP1	rs4971079	0.674	1.04E-05	0.661385	3.82E-07	0.98461
ICAM4	rs3093029	1.22	0.000483	1.300892	2.89E-08	0.693603
C17orf97	rs2857657	-0.747	1.63E-06	0.290567	0.142846	0.729101
EPHB4	rs7784933	-0.412	0.0477	-0.64414	4.86E-05	0.733902
MCCC1	rs2314737	-0.416	0.0414	-0.7201	2.08E-05	0.758154
PTPRS	rs10417781	-0.446	7.04E-05	0.375342	0.07337	0.760943

VIL1	rs56191141	-0.49	0.012	-0.66284	6.74E-06	0.77063
SMDT1	rs133341	-0.735	0.00251	-0.7165	4.30E-07	0.773525
LINC01260	rs2299686	-0.553	0.00686	-0.57997	3.18E-05	0.781926
FLJ42351	rs4848286	0.798	0.000204	0.898136	3.36E-08	0.810656
WARS	rs11160582	0.644	0.00914	0.589059	2.43E-06	0.819379
NUF2	rs10799928	1.42	1.61E-11	0.168798	0.335059	0.825918
NT5C3B	rs2304494	1.06	0.00123	1.09787	4.35E-06	0.859337
LINC01089	rs7974348	-0.645	0.00126	-0.59437	3.27E-06	0.862302
FAM206A	rs2304779	-0.718	5.01E-05	-0.60913	0.000188	0.864452
GSDMB	rs11078926	-0.507	0.00589	-0.56385	9.90E-07	0.866853
PAQR6	rs2274226	0.901	3.61E-06	-0.91761	1.02E-13	0.882854
SMG5	rs2274226	1	1.46E-07	-1.13558	2.83E-22	0.884265
IL12RB1	rs8109496	-0.823	0.00731	-0.78696	2.68E-06	0.912236
SULT1A1	rs7191548	-0.487	0.00645	-0.61054	5.31E-07	0.930373
ITLN1	rs2039415	-0.397	0.0806	-0.52389	6.34E-05	0.931464
CBS	rs2124458	0.575	0.00569	0.587712	1.49E-05	0.936199

Supplementary Table 5 – Colocalization analyses: Posterior Probabilities for each Hypothesis

A. IBD GWAS (trait 2); IBD eQTL (trait 1)									**D. IBD GWAS (trait 2); JIA eQTL (trait 1)**						
Gene	**rsID**	**H0**	**H1**	**H2**	**H3**	**H4**			**Gene**	**rsID**	**H0**	**H1**	**H2**	**H3**	**H4**
ABHD16B	rs6062504	0.00	0.00	0.92	0.01	0.08			ABHD16B	rs6062504	0.00	0.00	0.92	0.01	0.07
ACO2	rs727563	0.79	0.00	0.19	0.00	0.02			ACO2	rs727563	0.79	0.00	0.19	0.00	0.02
ACTR1A	rs3740415	0.01	0.00	0.91	0.00	0.07			ACTR1A	rs3740415	0.01	0.00	0.92	0.00	0.06
AGPAT2	rs13300218	0.00	0.00	0.91	0.01	0.08			AGPAT2	rs13300218	0.00	0.00	0.92	0.01	0.07
APOBEC3G	rs12627970	0.00	0.00	0.91	0.01	0.08			APOBEC3G	rs12627970	0.00	0.00	0.92	0.01	0.08
APOBEC3H	rs12627970	0.00	0.00	0.87	0.01	0.12			APOBEC3H	rs12627970	0.00	0.00	0.85	0.01	0.14
ARFRP1	rs6062504	0.00	0.00	0.92	0.01	0.08			ARFRP1	rs6062504	0.00	0.00	0.92	0.01	0.07
ARL3	rs3740415	0.01	0.00	0.82	0.00	0.17			ARL3	rs3740415	0.01	0.00	0.77	0.00	0.21
ATF4	rs12627970	0.00	0.00	0.90	0.01	0.09			ATF4	rs12627970	0.00	0.00	0.91	0.01	0.09
ATP9B	rs7236492	0.02	0.00	0.90	0.00	0.08			ATP9B	rs7236492	0.02	0.00	0.90	0.00	0.08
B4GALT7	rs4976646	0.00	0.00	0.92	0.00	0.08			B4GALT7	rs4976646	0.00	0.00	0.92	0.00	0.08
BANK1	rs13126505	0.00	0.00	0.90	0.01	0.09			BANK1	rs13126505	0.00	0.00	0.90	0.01	0.09
BAZ2B	rs4664304	0.04	0.00	0.88	0.00	0.08			BAZ2B	rs4664304	0.04	0.00	0.89	0.00	0.07
C16orf74	rs2361755	0.00	0.00	0.90	0.00	0.10			C16orf74	rs2361755	0.00	0.00	0.89	0.00	0.11
C17orf67	rs3853824	0.00	0.00	0.90	0.00	0.10			C17orf67	rs3853824	0.00	0.00	0.90	0.00	0.09
C22orf46	rs727563	0.79	0.00	0.19	0.00	0.02			C22orf46	rs727563	0.79	0.00	0.20	0.00	0.02
C5orf56	rs17622378	0.00	0.00	0.91	0.01	0.08			C5orf56	rs17622378	0.00	0.00	0.92	0.01	0.07
CACNA1I	rs12627970	0.00	0.00	0.90	0.00	0.10			CACNA1I	rs12627970	0.00	0.00	0.89	0.01	0.10
CARD9	rs13300218	0.00	0.00	0.92	0.01	0.07			CARD9	rs13300218	0.00	0.00	0.93	0.01	0.06
CBX7	rs12627970	0.00	0.00	0.90	0.02	0.08			CBX7	rs12627970	0.00	0.00	0.89	0.03	0.07
CD226	rs727088	0.00	0.00	0.86	0.01	0.13			CD226	rs727088	0.00	0.00	0.84	0.01	0.15
CD28	rs3116494	0.19	0.00	0.60	0.01	0.20			CD28	rs3116494	0.16	0.00	0.52	0.01	0.30
CD40	rs6074022	0.00	0.00	0.91	0.01	0.08			CD40	rs6074022	0.00	0.00	0.92	0.01	0.07
CDK12	rs12946510	0.00	0.00	0.92	0.01	0.07			CDK12	rs12946510	0.00	0.00	0.93	0.01	0.06
CEBPA	rs17694108	0.00	0.00	0.89	0.00	0.11			CEBPA	rs17694108	0.00	0.00	0.88	0.00	0.12
CEBPG	rs17694108	0.00	0.00	0.92	0.00	0.08			CEBPG	rs17694108	0.00	0.00	0.93	0.00	0.07

COASY	rs12942547	0.00	0.00	0.89	0.01	0.11			COASY	rs12942547	0.00	0.00	0.88	0.01	0.11
COIL	rs3853824	0.00	0.00	0.92	0.00	0.08			COIL	rs3853824	0.00	0.00	0.93	0.00	0.07
COX4I1	rs2361755	0.00	0.00	0.91	0.00	0.09			COX4I1	rs2361755	0.00	0.00	0.91	0.00	0.09
CPEB4	rs72810983	0.00	0.00	0.90	0.02	0.08			CPEB4	rs72810983	0.00	0.00	0.91	0.02	0.07
CRYZL1	rs2284553	0.01	0.00	0.90	0.00	0.09			CRYZL1	rs2284553	0.01	0.00	0.91	0.00	0.08
CTDP1	rs7236492	0.02	0.00	0.89	0.00	0.08			CTDP1	rs7236492	0.02	0.00	0.90	0.00	0.08
CTLA4	rs3116494	0.21	0.00	0.68	0.01	0.10			CTLA4	rs3116494	0.21	0.00	0.67	0.01	0.11
CTSA	rs6074022	0.00	0.00	0.91	0.01	0.08			CTSA	rs6074022	0.00	0.00	0.92	0.01	0.07
CTSZ	rs259964	0.00	0.00	0.89	0.00	0.11			CTSZ	rs259964	0.00	0.00	0.88	0.00	0.12
CUEDC2	rs3740415	0.01	0.00	0.91	0.00	0.07			CUEDC2	rs3740415	0.01	0.00	0.92	0.00	0.06
CYTH1	rs17736589	0.91	0.00	0.08	0.00	0.01			CYTH1	rs17736589	0.91	0.00	0.08	0.00	0.01
DBN1	rs4976646	0.00	0.00	0.90	0.00	0.10			DBN1	rs4976646	0.00	0.00	0.91	0.00	0.09
DCTN5	rs7404095	0.06	0.00	0.87	0.00	0.07			DCTN5	rs7404095	0.06	0.00	0.88	0.00	0.06
DDX41	rs4976646	0.00	0.00	0.91	0.00	0.09			DDX41	rs4976646	0.00	0.00	0.92	0.00	0.08
DENND1B	rs2488389	0.00	0.00	0.92	0.01	0.08			DENND1B	rs2488389	0.00	0.00	0.92	0.01	0.07
DESI1	rs727563	0.77	0.00	0.19	0.00	0.05			DESI1	rs727563	0.75	0.00	0.19	0.00	0.06
DGKE	rs3853824	0.00	0.00	0.92	0.00	0.08			DGKE	rs3853824	0.00	0.00	0.93	0.00	0.07
DNAH17	rs17736589	0.91	0.00	0.08	0.00	0.01			DNAH17	rs17736589	0.91	0.00	0.08	0.00	0.01
DNAJC5	rs6062504	0.00	0.00	0.92	0.01	0.08			DNAJC5	rs6062504	0.00	0.00	0.93	0.01	0.07
DNLZ	rs13300218	0.00	0.00	0.92	0.01	0.07			DNLZ	rs13300218	0.00	0.00	0.93	0.01	0.06
DOK3	rs4976646	0.00	0.00	0.93	0.00	0.07			DOK3	rs4976646	0.00	0.00	0.94	0.00	0.06
DONSON	rs2284553	0.01	0.00	0.91	0.00	0.08			DONSON	rs2284553	0.01	0.00	0.92	0.00	0.07
DPH5	rs11583043	0.24	0.00	0.66	0.01	0.09			DPH5	rs11583043	0.24	0.00	0.65	0.01	0.11
ECHDC1	rs2503322	0.50	0.00	0.46	0.00	0.04			ECHDC1	rs2503322	0.50	0.00	0.46	0.00	0.03
EGFL7	rs13300218	0.00	0.00	0.77	0.01	0.22			EGFL7	rs13300218	0.00	0.00	0.69	0.01	0.30
ELOVL3	rs3740415	0.17	0.00	0.75	0.00	0.08			ELOVL3	rs3740415	0.17	0.00	0.75	0.00	0.08
EMC8	rs2361755	0.00	0.00	0.89	0.00	0.10			EMC8	rs2361755	0.00	0.00	0.88	0.00	0.12
ERBB2	rs12946510	0.00	0.00	0.92	0.01	0.07			ERBB2	rs12946510	0.00	0.00	0.93	0.01	0.06
EXTL2	rs11583043	0.23	0.00	0.63	0.01	0.13			EXTL2	rs11583043	0.22	0.00	0.60	0.01	0.17
F12	rs4976646	0.00	0.00	0.91	0.00	0.09			F12	rs4976646	0.00	0.00	0.91	0.00	0.09
FAM134C	rs12942547	0.00	0.00	0.90	0.01	0.09			FAM134C	rs12942547	0.00	0.00	0.91	0.01	0.09
FAM193B	rs4976646	0.00	0.00	0.81	0.00	0.19			FAM193B	rs4976646	0.00	0.00	0.75	0.00	0.25
FAM69B	rs13300218	0.00	0.00	0.87	0.00	0.12			FAM69B	rs13300218	0.00	0.00	0.86	0.00	0.13

FBXL15	rs3740415	0.01	0.00	0.91	0.00	0.07			FBXL15	rs3740415	0.01	0.00	0.92	0.00	0.06
FYN	rs3851228	0.00	0.00	0.90	0.01	0.09			FYN	rs3851228	0.00	0.00	0.91	0.01	0.08
GART	rs2284553	0.01	0.00	0.91	0.00	0.07			GART	rs2284553	0.01	0.00	0.92	0.00	0.07
GBF1	rs3740415	0.01	0.00	0.87	0.00	0.12			GBF1	rs3740415	0.01	0.00	0.86	0.00	0.13
GHDC	rs12942547	0.00	0.00	0.87	0.01	0.13			GHDC	rs12942547	0.00	0.00	0.85	0.01	0.14
GLS	rs1517352	0.00	0.00	0.92	0.01	0.08			GLS	rs1517352	0.00	0.00	0.93	0.01	0.07
GMEB2	rs6062504	0.00	0.00	0.92	0.01	0.08			GMEB2	rs6062504	0.00	0.00	0.93	0.01	0.07
GPATCH1	rs17694108	0.00	0.00	0.92	0.00	0.08			GPATCH1	rs17694108	0.00	0.00	0.93	0.00	0.07
GPSM1	rs13300218	0.00	0.00	0.88	0.01	0.11			GPSM1	rs13300218	0.00	0.00	0.88	0.01	0.11
GRK6	rs4976646	0.00	0.00	0.91	0.00	0.09			GRK6	rs4976646	0.00	0.00	0.91	0.00	0.09
GSDMB	rs12946510	0.00	0.00	0.22	0.01	0.76			GSDMB	rs12946510	0.00	0.00	0.07	0.01	0.92
HELZ2	rs6062504	0.00	0.00	0.91	0.01	0.08			HELZ2	rs6062504	0.00	0.00	0.92	0.01	0.07
HSD17B1	rs12942547	0.00	0.00	0.86	0.01	0.13			HSD17B1	rs12942547	0.00	0.00	0.84	0.01	0.15
ICOS	rs3116494	0.20	0.00	0.63	0.01	0.16			ICOS	rs3116494	0.18	0.00	0.58	0.01	0.23
IFNAR1	rs2284553	0.01	0.00	0.91	0.00	0.08			IFNAR1	rs2284553	0.01	0.00	0.92	0.00	0.07
IFNAR2	rs2284553	0.01	0.00	0.91	0.00	0.08			IFNAR2	rs2284553	0.01	0.00	0.92	0.00	0.07
IFNGR2	rs2284553	0.01	0.00	0.91	0.00	0.08			IFNGR2	rs2284553	0.01	0.00	0.92	0.00	0.07
IKZF3	rs12946510	0.00	0.00	0.91	0.01	0.08			IKZF3	rs12946510	0.00	0.00	0.91	0.01	0.08
IL10RB	rs2284553	0.01	0.00	0.91	0.00	0.08			IL10RB	rs2284553	0.01	0.00	0.92	0.00	0.07
INPP5E	rs13300218	0.00	0.00	0.89	0.01	0.10			INPP5E	rs13300218	0.00	0.00	0.89	0.01	0.10
IRF1	rs17622378	0.00	0.00	0.90	0.02	0.08			IRF1	rs17622378	0.00	0.00	0.91	0.02	0.07
IRF8	rs2361755	0.00	0.00	0.91	0.00	0.09			IRF8	rs2361755	0.00	0.00	0.91	0.00	0.09
ITSN1	rs2284553	0.01	0.00	0.91	0.00	0.08			ITSN1	rs2284553	0.01	0.00	0.92	0.00	0.07
KIF3A	rs17622378	0.00	0.00	0.91	0.01	0.09			KIF3A	rs17622378	0.00	0.00	0.91	0.01	0.08
L3MBTL2	rs727563	0.79	0.00	0.19	0.00	0.02			L3MBTL2	rs727563	0.79	0.00	0.20	0.00	0.01
LCN10	rs13300218	0.00	0.00	0.92	0.00	0.08			LCN10	rs13300218	0.00	0.00	0.93	0.00	0.07
LGALS3BP	rs17736589	0.99	0.00	0.01	0.00	0.00			LGALS3BP	rs17736589	0.99	0.00	0.01	0.00	0.00
LIME1	rs6062504	0.00	0.00	0.90	0.01	0.09			LIME1	rs6062504	0.00	0.00	0.90	0.01	0.09
LINC01573	rs13300218	0.00	0.00	0.86	0.01	0.13			LINC01573	rs13300218	0.00	0.00	0.84	0.01	0.15
LITAF	rs11641184	0.00	0.00	0.89	0.00	0.11			LITAF	rs11641184	0.00	0.00	0.89	0.00	0.11
LMAN2	rs4976646	0.00	0.00	0.91	0.00	0.09			LMAN2	rs4976646	0.00	0.00	0.92	0.00	0.08
LOC101927131	rs11641184	0.00	0.00	0.89	0.00	0.10			LOC101927131	rs11641184	0.00	0.00	0.89	0.00	0.11
LOC101928370	rs11583043	0.36	0.00	0.57	0.00	0.07			LOC101928370	rs11583043	0.35	0.00	0.56	0.00	0.08

LOC102606465	rs11583043	0.24	0.00	0.67	0.01	0.08			LOC102606465	rs11583043	0.24	0.00	0.66	0.01	0.08
LRP3	rs17694108	0.00	0.00	0.89	0.00	0.11			LRP3	rs17694108	0.00	0.00	0.88	0.00	0.12
MANBA	rs3774937	0.94	0.00	0.06	0.00	0.01			MANBA	rs3774937	0.94	0.00	0.06	0.00	0.01
MEI1	rs727563	0.79	0.00	0.19	0.00	0.02			MEI1	rs727563	0.79	0.00	0.19	0.00	0.02
MGAT3	rs12627970	0.00	0.00	0.90	0.02	0.08			MGAT3	rs12627970	0.00	0.00	0.84	0.09	0.07
MIEF1	rs12627970	0.00	0.00	0.90	0.01	0.08			MIEF1	rs12627970	0.00	0.00	0.90	0.02	0.08
MIEN1	rs12946510	0.00	0.00	0.92	0.01	0.08			MIEN1	rs12946510	0.00	0.00	0.93	0.01	0.07
MIR647	rs6062504	0.00	0.00	0.75	0.00	0.24			MIR647	rs6062504	0.00	0.00	0.64	0.00	0.36
MLX	rs12942547	0.00	0.00	0.67	0.01	0.32			MLX	rs12942547	0.00	0.00	0.51	0.01	0.48
MMP9	rs6074022	0.00	0.00	0.90	0.01	0.09			MMP9	rs6074022	0.00	0.00	0.90	0.01	0.09
MXD3	rs4976646	0.00	0.00	0.93	0.00	0.07			MXD3	rs4976646	0.00	0.00	0.94	0.00	0.06
NAGLU	rs12942547	0.00	0.00	0.91	0.01	0.08			NAGLU	rs12942547	0.00	0.00	0.92	0.01	0.08
NCOA5	rs6074022	0.00	0.00	0.89	0.01	0.10			NCOA5	rs6074022	0.00	0.00	0.89	0.01	0.10
NFATC1	rs7236492	0.02	0.00	0.89	0.00	0.09			NFATC1	rs7236492	0.02	0.00	0.89	0.00	0.08
NFKB1	rs3774937	0.94	0.00	0.06	0.00	0.01			NFKB1	rs3774937	0.94	0.00	0.06	0.00	0.00
NFKB2	rs3740415	0.01	0.00	0.91	0.00	0.08			NFKB2	rs3740415	0.01	0.00	0.91	0.00	0.07
NHP2L1	rs727563	0.79	0.00	0.19	0.00	0.02			NHP2L1	rs727563	0.79	0.00	0.20	0.00	0.01
NOG	rs3853824	0.00	0.00	0.91	0.00	0.09			NOG	rs3853824	0.00	0.00	0.92	0.00	0.08
NOTCH1	rs13300218	0.00	0.00	0.90	0.01	0.09			NOTCH1	rs13300218	0.00	0.00	0.90	0.01	0.09
NSD1	rs4976646	0.00	0.00	0.92	0.00	0.08			NSD1	rs4976646	0.00	0.00	0.93	0.00	0.07
ORMDL3	rs12946510	0.00	0.00	0.67	0.01	0.32			ORMDL3	rs12946510	0.00	0.00	0.52	0.01	0.46
PALB2	rs7404095	0.06	0.00	0.86	0.00	0.08			PALB2	rs7404095	0.06	0.00	0.86	0.00	0.08
PCIF1	rs6074022	0.00	0.00	0.90	0.01	0.09			PCIF1	rs6074022	0.00	0.00	0.90	0.01	0.09
PDGFB	rs12627970	0.00	0.00	0.90	0.01	0.09			PDGFB	rs12627970	0.00	0.00	0.90	0.01	0.09
PDLIM7	rs4976646	0.00	0.00	0.91	0.00	0.09			PDLIM7	rs4976646	0.00	0.00	0.92	0.00	0.08
PEPD	rs17694108	0.00	0.00	0.86	0.00	0.14			PEPD	rs17694108	0.00	0.00	0.82	0.00	0.18
PGAP3	rs12946510	0.00	0.00	0.92	0.01	0.08			PGAP3	rs12946510	0.00	0.00	0.93	0.01	0.07
PHF5A	rs727563	0.78	0.00	0.19	0.00	0.02			PHF5A	rs727563	0.78	0.00	0.19	0.00	0.02
PLCL1	rs1440088	1.00	0.00	0.00	0.00	0.00			PLCL1	rs1440088	1.00	0.00	0.00	0.00	0.00
PLK1	rs7404095	0.06	0.00	0.85	0.00	0.09			PLK1	rs7404095	0.06	0.00	0.86	0.00	0.09
PLTP	rs6074022	0.00	0.00	0.82	0.01	0.17			PLTP	rs6074022	0.00	0.00	0.76	0.01	0.23
PMM1	rs727563	0.79	0.00	0.19	0.00	0.02			PMM1	rs727563	0.79	0.00	0.19	0.00	0.02
PMPCA	rs13300218	0.00	0.00	0.92	0.01	0.07			PMPCA	rs13300218	0.00	0.00	0.93	0.01	0.06

POLR3H	rs727563	0.79	0.00	0.19	0.00	0.02			POLR3H	rs727563	0.79	0.00	0.20	0.00	0.02
PPA2	rs2189234	0.00	0.00	0.93	0.00	0.07			PPA2	rs2189234	0.00	0.00	0.93	0.00	0.06
PPDPF	rs6062504	0.00	0.00	0.91	0.01	0.09			PPDPF	rs6062504	0.00	0.00	0.91	0.01	0.08
PRELID1	rs4976646	0.00	0.00	0.91	0.00	0.09			PRELID1	rs4976646	0.00	0.00	0.92	0.00	0.08
PRKCB	rs7404095	0.03	0.00	0.46	0.00	0.50			PRKCB	rs7404095	0.02	0.00	0.23	0.00	0.75
PRR7	rs4976646	0.00	0.00	0.92	0.00	0.08			PRR7	rs4976646	0.00	0.00	0.93	0.00	0.07
PSD	rs3740415	0.01	0.00	0.90	0.00	0.09			PSD	rs3740415	0.01	0.00	0.91	0.00	0.08
PSMC3IP	rs12942547	0.00	0.00	0.92	0.01	0.07			PSMC3IP	rs12942547	0.00	0.00	0.93	0.01	0.07
PSMD3	rs12946510	0.00	0.00	0.91	0.00	0.08			PSMD3	rs12946510	0.00	0.00	0.92	0.00	0.08
PTK6	rs6062504	0.00	0.00	0.90	0.01	0.09			PTK6	rs6062504	0.00	0.00	0.90	0.01	0.09
PTPRK	rs13204742	0.00	0.00	0.88	0.01	0.11			PTPRK	rs13204742	0.00	0.00	0.87	0.01	0.12
RAB24	rs4976646	0.00	0.00	0.92	0.00	0.08			RAB24	rs4976646	0.00	0.00	0.92	0.00	0.08
RAB5C	rs12942547	0.00	0.00	0.92	0.00	0.08			RAB5C	rs12942547	0.00	0.00	0.93	0.00	0.07
RAD50	rs17622378	0.00	0.00	0.90	0.01	0.09			RAD50	rs17622378	0.00	0.00	0.91	0.01	0.08
RANGAP1	rs727563	0.79	0.00	0.19	0.00	0.02			RANGAP1	rs727563	0.79	0.00	0.20	0.00	0.02
REV3L	rs3851228	0.00	0.00	0.90	0.01	0.09			REV3L	rs3851228	0.00	0.00	0.90	0.01	0.08
RGS14	rs4976646	0.00	0.00	0.87	0.00	0.13			RGS14	rs4976646	0.00	0.00	0.85	0.00	0.15
RNF145	rs56167332	0.00	0.00	0.88	0.03	0.10			RNF145	rs56167332	0.00	0.00	0.87	0.03	0.10
RNF146	rs2503322	0.50	0.00	0.46	0.00	0.04			RNF146	rs2503322	0.50	0.00	0.46	0.00	0.03
RPL3	rs12627970	0.00	0.00	0.88	0.03	0.09			RPL3	rs12627970	0.00	0.00	0.84	0.08	0.08
RPS19BP1	rs12627970	0.00	0.00	0.90	0.01	0.09			RPS19BP1	rs12627970	0.00	0.00	0.91	0.01	0.08
RSL1D1	rs11641184	0.00	0.00	0.89	0.00	0.11			RSL1D1	rs11641184	0.00	0.00	0.88	0.00	0.12
RTTN	rs727088	0.00	0.00	0.91	0.01	0.08			RTTN	rs727088	0.00	0.00	0.91	0.01	0.08
S1PR1	rs11583043	0.35	0.00	0.56	0.00	0.08			S1PR1	rs11583043	0.35	0.00	0.55	0.00	0.09
SCPEP1	rs3853824	0.00	0.00	0.91	0.00	0.09			SCPEP1	rs3853824	0.00	0.00	0.91	0.00	0.09
SDCCAG3	rs13300218	0.00	0.00	0.88	0.01	0.10			SDCCAG3	rs13300218	0.00	0.00	0.88	0.01	0.11
SEC16A	rs13300218	0.00	0.00	0.92	0.01	0.07			SEC16A	rs13300218	0.00	0.00	0.93	0.01	0.06
SFXN2	rs3740415	0.01	0.00	0.88	0.00	0.11			SFXN2	rs3740415	0.01	0.00	0.88	0.00	0.11
SLC22A4	rs17622378	0.00	0.00	0.91	0.01	0.07			SLC22A4	rs17622378	0.00	0.00	0.92	0.01	0.06
SLC22A5	rs17622378	0.00	0.00	0.00	0.98	0.02			SLC22A5	rs17622378	0.00	0.00	0.00	1.00	0.00
SLC2A4RG	rs6062504	0.00	0.00	0.90	0.01	0.09			SLC2A4RG	rs6062504	0.00	0.00	0.91	0.01	0.08
SLC30A7	rs11583043	0.24	0.00	0.66	0.01	0.09			SLC30A7	rs11583043	0.23	0.00	0.64	0.01	0.11
SLC35C2	rs6074022	0.00	0.00	0.91	0.00	0.09			SLC35C2	rs6074022	0.00	0.00	0.91	0.00	0.08

SLC39A8	rs3774937	0.94	0.00	0.06	0.00	0.01			SLC39A8	rs3774937	0.94	0.00	0.06	0.00	0.01
SNAPC4	rs13300218	0.00	0.00	0.90	0.01	0.09			SNAPC4	rs13300218	0.00	0.00	0.91	0.01	0.09
SNN	rs11641184	0.00	0.00	0.86	0.00	0.14			SNN	rs11641184	0.00	0.00	0.84	0.00	0.16
STARD3	rs12946510	0.00	0.00	0.84	0.01	0.15			STARD3	rs12946510	0.00	0.00	0.81	0.01	0.18
STAT1	rs1517352	0.00	0.00	0.90	0.01	0.09			STAT1	rs1517352	0.00	0.00	0.90	0.01	0.09
STAT3	rs12942547	0.00	0.00	0.84	0.01	0.15			STAT3	rs12942547	0.00	0.00	0.80	0.01	0.19
STAT4	rs1517352	0.00	0.00	0.92	0.01	0.07			STAT4	rs1517352	0.00	0.00	0.93	0.01	0.06
STAT5A	rs12942547	0.00	0.00	0.92	0.01	0.08			STAT5A	rs12942547	0.00	0.00	0.93	0.01	0.07
STAT5B	rs12942547	0.00	0.00	0.91	0.01	0.09			STAT5B	rs12942547	0.00	0.00	0.91	0.01	0.08
STMN3	rs6062504	0.00	0.00	0.92	0.01	0.08			STMN3	rs6062504	0.00	0.00	0.92	0.01	0.07
SUFU	rs3740415	0.01	0.00	0.91	0.00	0.08			SUFU	rs3740415	0.01	0.00	0.92	0.00	0.07
SYNGR1	rs12627970	0.00	0.00	0.91	0.01	0.08			SYNGR1	rs12627970	0.00	0.00	0.92	0.01	0.07
TCAP	rs12946510	0.00	0.00	0.92	0.01	0.07			TCAP	rs12946510	0.00	0.00	0.93	0.01	0.06
TEF	rs727563	0.79	0.00	0.19	0.00	0.02			TEF	rs727563	0.79	0.00	0.19	0.00	0.02
TET2	rs2189234	0.00	0.00	0.92	0.00	0.08			TET2	rs2189234	0.00	0.00	0.92	0.00	0.08
THEMIS	rs13204742	0.00	0.00	0.88	0.01	0.11			THEMIS	rs13204742	0.00	0.00	0.86	0.01	0.12
TIMP2	rs17736589	0.91	0.00	0.08	0.00	0.01			TIMP2	rs17736589	0.91	0.00	0.08	0.00	0.01
TMED9	rs4976646	0.00	0.00	0.93	0.00	0.07			TMED9	rs4976646	0.00	0.00	0.94	0.00	0.06
TMEM180	rs3740415	0.00	0.00	0.26	0.00	0.74			TMEM180	rs3740415	0.00	0.00	0.08	0.00	0.92
TMEM50B	rs2284553	0.01	0.00	0.91	0.00	0.08			TMEM50B	rs2284553	0.01	0.00	0.92	0.00	0.07
TOB2	rs727563	0.79	0.00	0.19	0.00	0.02			TOB2	rs727563	0.79	0.00	0.20	0.00	0.01
TPD52L2	rs6062504	0.00	0.00	0.91	0.01	0.08			TPD52L2	rs6062504	0.00	0.00	0.92	0.01	0.07
TRAF3IP2	rs3851228	0.00	0.00	0.90	0.01	0.09			TRAF3IP2	rs3851228	0.00	0.00	0.90	0.02	0.08
TRIM8	rs3740415	0.01	0.00	0.90	0.00	0.09			TRIM8	rs3740415	0.01	0.00	0.91	0.00	0.08
TUBB1	rs259964	0.00	0.00	0.91	0.00	0.09			TUBB1	rs259964	0.00	0.00	0.91	0.00	0.09
TUBG1	rs12942547	0.00	0.00	0.91	0.00	0.08			TUBG1	rs12942547	0.00	0.00	0.92	0.00	0.08
TXNDC11	rs11641184	0.00	0.00	0.87	0.00	0.13			TXNDC11	rs11641184	0.00	0.00	0.85	0.00	0.15
UBLCP1	rs56167332	0.00	0.00	0.89	0.03	0.08			UBLCP1	rs56167332	0.00	0.00	0.89	0.03	0.07
UCKL1	rs6062504	0.00	0.00	0.91	0.00	0.09			UCKL1	rs6062504	0.00	0.00	0.91	0.00	0.09
USP36	rs17736589	0.91	0.00	0.08	0.00	0.01			USP36	rs17736589	0.91	0.00	0.08	0.00	0.01
XRCC6	rs727563	0.79	0.00	0.19	0.00	0.02			XRCC6	rs727563	0.79	0.00	0.20	0.00	0.01
ZBTB46	rs6062504	0.00	0.00	0.91	0.01	0.08			ZBTB46	rs6062504	0.00	0.00	0.92	0.01	0.07
ZC3H7A	rs11641184	0.00	0.00	0.90	0.00	0.10			ZC3H7A	rs11641184	0.00	0.00	0.90	0.00	0.10

ZC3H7B	rs727563	0.79	0.00	0.19	0.00	0.02			ZC3H7B	rs727563	0.79	0.00	0.20	0.00	0.02
ZGPAT	rs6062504	0.00	0.00	0.92	0.01	0.08			ZGPAT	rs6062504	0.00	0.00	0.93	0.01	0.07
ZNF335	rs6074022	0.00	0.00	0.91	0.01	0.08			ZNF335	rs6074022	0.00	0.00	0.91	0.01	0.08
ZNF512B	rs6062504	0.00	0.00	0.92	0.00	0.08			ZNF512B	rs6062504	0.00	0.00	0.93	0.00	0.07
ZNF831	rs259964	0.00	0.00	0.92	0.00	0.08			ZNF831	rs259964	0.00	0.00	0.93	0.00	0.07
ZSWIM1	rs6074022	0.00	0.00	0.91	0.01	0.08			ZSWIM1	rs6074022	0.00	0.00	0.92	0.01	0.08
ZSWIM3	rs6074022	0.00	0.00	0.91	0.01	0.09			ZSWIM3	rs6074022	0.00	0.00	0.91	0.01	0.08
B. JIA GWAS (trait 2); JIA eQTL (trait 1)									**E. JIA GWAS (trait 2); IBD eQTL (trait 1)**						
Gene	**rsID**	**H0**	**H1**	**H2**	**H3**	**H4**			**Gene**	**rsID**	**H0**	**H1**	**H2**	**H3**	**H4**
ADPRH	rs4688013	0.10	0.12	0.34	0.41	0.03			ADPRH	rs4688013	0.09	0.13	0.32	0.43	0.03
ARHGAP31	rs4688013	0.11	0.12	0.35	0.39	0.03			ARHGAP31	rs4688013	0.10	0.13	0.33	0.42	0.03
CEP192	rs2847293	0.05	0.05	0.45	0.40	0.04			CEP192	rs2847293	0.04	0.03	0.49	0.39	0.05
CEP76	rs2847293	0.05	0.04	0.47	0.39	0.04			CEP76	rs2847293	0.04	0.03	0.51	0.38	0.04
CHST10	rs6740838	0.78	0.09	0.11	0.01	0.01			CHST10	rs6740838	0.78	0.08	0.12	0.01	0.01
CLEC16A	rs66718203	0.00	0.00	0.24	0.73	0.02			CLEC16A	rs66718203	0.00	0.00	0.21	0.76	0.03
COX17	rs4688013	0.11	0.12	0.36	0.38	0.03			COX17	rs4688013	0.10	0.12	0.33	0.41	0.04
KIAA1109	rs1479924	0.42	0.05	0.43	0.05	0.04									
LITAF	rs66718203	0.00	0.01	0.30	0.66	0.03			LITAF	rs66718203	0.00	0.00	0.30	0.67	0.03
LOC100996324	rs2847293	0.05	0.04	0.45	0.39	0.06			LOC100996324	rs2847293	0.02	0.04	0.22	0.52	0.20
LOC101927131	rs66718203	0.00	0.00	0.24	0.73	0.02			LOC101927131	rs66718203	0.00	0.00	0.23	0.75	0.02
POGLUT1	rs4688013	0.10	0.12	0.33	0.42	0.04			POGLUT1	rs4688013	0.08	0.14	0.28	0.47	0.03
POPDC2	rs4688013	0.10	0.13	0.33	0.42	0.03			POPDC2	rs4688013	0.09	0.14	0.29	0.46	0.03
PSMG2	rs2847293	0.06	0.04	0.48	0.38	0.04			PSMG2	rs2847293	0.04	0.03	0.51	0.38	0.04
PTPN2	rs2847293	0.06	0.04	0.49	0.37	0.04			PTPN2	rs2847293	0.04	0.03	0.50	0.39	0.04
RMI2	rs66718203	0.00	0.00	0.10	0.88	0.01			RMI2	rs66718203	0.00	0.00	0.04	0.96	0.00
SEH1L	rs2847293	0.05	0.04	0.47	0.39	0.04			SEH1L	rs2847293	0.04	0.03	0.51	0.38	0.04
SOCS1	rs66718203	0.00	0.00	0.23	0.74	0.02			SOCS1	rs66718203	0.00	0.00	0.22	0.76	0.02
SPIRE1	rs2847293	0.05	0.04	0.46	0.38	0.06			SPIRE1	rs2847293	0.04	0.03	0.47	0.41	0.04
TIMMDC1	rs4688013	0.11	0.12	0.35	0.39	0.03			TIMMDC1	rs4688013	0.10	0.12	0.33	0.42	0.03
TMEM39A	rs4688013	0.11	0.12	0.35	0.39	0.03			TMEM39A	rs4688013	0.10	0.13	0.33	0.42	0.03

C. RA GWAS (trait 2); JIA eQTL (trait 1)									**F. RA GWAS (trait 2); IBD eQTL (trait 1)**						
Gene	**rsID**	**H0**	**H1**	**H2**	**H3**	**H4**			**Gene**	**rsID**	**H0**	**H1**	**H2**	**H3**	**H4**
									ALS2CR12	rs6715284	0.06	0.01	0.77	0.07	0.09
ARHGAP30	rs4656942	0.99	0.01	0.00	0.00	0.00									
									ATG5	rs9372120	0.43	0.01	0.50	0.01	0.05
									BUB1	rs6732565	0.87	0.01	0.11	0.00	0.01
C16orf74	rs13330176	0.01	0.00	0.92	0.00	0.07			C16orf74	rs13330176	0.01	0.00	0.91	0.01	0.07
C5orf30	rs2561477	0.00	0.00	0.92	0.01	0.08			C5orf30	rs2561477	0.00	0.00	0.92	0.01	0.07
									CASP10	rs6715284	0.07	0.00	0.83	0.02	0.07
									CASP8	rs6715284	0.06	0.00	0.81	0.02	0.10
CCR6	rs1571878	0.00	0.00	0.92	0.01	0.08			CCR6	rs1571878	0.00	0.00	0.53	0.03	0.44
CD2	rs624988	0.01	0.00	0.90	0.01	0.08			CD2	rs624988	0.01	0.00	0.91	0.01	0.07
CD244	rs4656942	0.99	0.01	0.00	0.00	0.00									
CD40	rs4239702	0.00	0.00	0.91	0.01	0.08			CD40	rs4239702	0.00	0.00	0.92	0.01	0.07
CD48	rs4656942	0.99	0.01	0.00	0.00	0.00									
CD58	rs624988	0.01	0.00	0.91	0.01	0.07			CD58	rs624988	0.01	0.00	0.91	0.01	0.07
CDK12	rs1877030	0.01	0.00	0.86	0.01	0.12			CDK12	rs1877030	0.01	0.00	0.90	0.02	0.07
									CEP192	rs8083786	0.00	0.00	0.86	0.07	0.06
									CEP76	rs8083786	0.00	0.00	0.92	0.02	0.07
									CFAP126	rs72717009	0.00	0.00	0.76	0.16	0.08
									CFLAR	rs6715284	0.06	0.00	0.82	0.02	0.09
COX4I1	rs13330176	0.01	0.00	0.91	0.00	0.08			COX4I1	rs13330176	0.01	0.00	0.90	0.02	0.07
CTSA	rs4239702	0.00	0.00	0.92	0.01	0.07			CTSA	rs4239702	0.00	0.00	0.92	0.01	0.07
ELMO2	rs4239702	0.00	0.00	0.92	0.00	0.08			ELMO2	rs4239702	0.00	0.00	0.92	0.01	0.07
EMC8	rs13330176	0.01	0.00	0.91	0.00	0.08			EMC8	rs13330176	0.01	0.00	0.90	0.02	0.07
ERBB2	rs1877030	0.01	0.00	0.90	0.01	0.08			ERBB2	rs1877030	0.01	0.00	0.91	0.02	0.06
F11R	rs4656942	0.99	0.01	0.00	0.00	0.00									
									FAM126B	rs6715284	0.12	0.00	0.79	0.02	0.07
FBXL20	rs1877030	0.05	0.00	0.86	0.01	0.08			FBXL20	rs1877030	0.03	0.00	0.87	0.01	0.08
									FCER1G	rs72717009	0.03	0.00	0.78	0.11	0.07
									FCGR2A	rs72717009	0.00	0.00	0.89	0.03	0.08

									FCGR2B	rs72717009	0.00	0.00	0.60	0.02	0.38
									FCGR2C	rs72717009	0.00	0.00	0.88	0.03	0.09
									FCGR3A	rs72717009	0.00	0.00	0.85	0.08	0.08
									FCGR3B	rs72717009	0.00	0.00	0.89	0.02	0.09
FGFR1OP	rs1571878	0.00	0.00	0.90	0.01	0.09			FGFR1OP	rs1571878	0.00	0.00	0.87	0.06	0.07
GIN1	rs2561477	0.00	0.00	0.92	0.01	0.08			GIN1	rs2561477	0.00	0.00	0.84	0.02	0.14
GLS	rs11889341	0.00	0.00	0.91	0.01	0.08			GLS	rs11889341	0.00	0.00	0.83	0.01	0.15
HIC2	rs11089637	0.11	0.00	0.80	0.00	0.08			HIC2	rs11089637	0.03	0.00	0.88	0.01	0.08
									HSPA6	rs72717009	0.00	0.00	0.86	0.05	0.10
									HSPA7	rs72717009	0.00	0.00	0.80	0.02	0.18
IKZF3	rs1877030	0.01	0.00	0.86	0.01	0.11			IKZF3	rs1877030	0.01	0.00	0.91	0.01	0.07
IRF8	rs13330176	0.01	0.00	0.91	0.00	0.08			IRF8	rs13330176	0.01	0.00	0.90	0.02	0.07
ITLN1	rs4656942	0.99	0.01	0.00	0.00	0.00									
									KIAA1109	rs45475795	0.00	0.00	0.89	0.02	0.10
									LOC100996324	rs8083786	0.00	0.00	0.77	0.15	0.08
									LOC101928673	rs2105325	0.00	0.00	0.90	0.02	0.08
LY9	rs4656942	0.98	0.01	0.00	0.00	0.00									
MAPK1	rs11089637	0.11	0.00	0.81	0.00	0.07			MAPK1	rs11089637	0.03	0.00	0.85	0.02	0.10
MED1	rs1877030	0.01	0.00	0.77	0.02	0.20			MED1	rs1877030	0.01	0.00	0.90	0.02	0.07
MIEN1	rs1877030	0.01	0.00	0.87	0.01	0.11			MIEN1	rs1877030	0.01	0.00	0.90	0.02	0.07
MMP9	rs4239702	0.00	0.00	0.92	0.01	0.07			MMP9	rs4239702	0.00	0.00	0.89	0.01	0.10
									MPZ	rs72717009	0.00	0.00	0.86	0.03	0.11
NCOA5	rs4239702	0.00	0.00	0.91	0.01	0.08			NCOA5	rs4239702	0.00	0.00	0.88	0.01	0.10
									NDUFB3	rs6715284	0.09	0.00	0.80	0.02	0.09
									NDUFS2	rs72717009	0.03	0.00	0.83	0.02	0.12
PAM	rs2561477	0.00	0.00	0.91	0.01	0.08			PAM	rs2561477	0.00	0.00	0.03	0.80	0.18
PCIF1	rs4239702	0.00	0.00	0.92	0.01	0.08			PCIF1	rs4239702	0.00	0.00	0.90	0.01	0.09
									PCP4L1	rs72717009	0.00	0.00	0.68	0.06	0.25
PFDN2	rs4656942	0.99	0.01	0.00	0.00	0.00									
PGAP3	rs1877030	0.01	0.00	0.86	0.01	0.12			PGAP3	rs1877030	0.01	0.00	0.89	0.02	0.09
PI4KAP2	rs11089637	0.11	0.00	0.80	0.00	0.09			PI4KAP2	rs11089637	0.03	0.00	0.85	0.02	0.11
PLTP	rs4239702	0.00	0.00	0.73	0.01	0.26			PLTP	rs4239702	0.00	0.00	0.72	0.02	0.26
PPIL2	rs11089637	0.11	0.00	0.81	0.00	0.08			PPIL2	rs11089637	0.03	0.00	0.86	0.03	0.08

PPIP5K2	rs2561477	0.00	0.00	0.89	0.01	0.09			PPIP5K2	rs2561477	0.00	0.00	0.92	0.01	0.07
									PRDM1	rs9372120	0.41	0.01	0.48	0.01	0.08
									PRDX6	rs2105325	0.00	0.00	0.89	0.03	0.08
									PSMG2	rs8083786	0.00	0.00	0.91	0.02	0.07
PTGFRN	rs624988	0.01	0.00	0.89	0.01	0.09			PTGFRN	rs624988	0.01	0.00	0.91	0.01	0.07
									PTPN2	rs8083786	0.00	0.00	0.91	0.02	0.07
PTPRC	rs17668708	0.38	0.00	0.54	0.00	0.08			PTPRC	rs17668708	0.15	0.00	0.77	0.01	0.06
RNASET2	rs1571878	0.00	0.00	0.87	0.01	0.12			RNASET2	rs1571878	0.00	0.00	0.00	0.99	0.01
									RSPH3	rs2451258	0.00	0.00	0.88	0.03	0.09
SDF2L1	rs11089637	0.11	0.00	0.81	0.00	0.08			SDF2L1	rs11089637	0.03	0.00	0.82	0.02	0.14
									SDHC	rs72717009	0.00	0.00	0.89	0.03	0.08
									SEH1L	rs8083786	0.00	0.00	0.91	0.02	0.07
SLAMF1	rs4656942	0.99	0.01	0.00	0.00	0.00									
SLAMF7	rs4656942	0.98	0.01	0.00	0.00	0.00									
SLC35C2	rs4239702	0.00	0.00	0.92	0.01	0.07			SLC35C2	rs4239702	0.00	0.00	0.91	0.01	0.08
									SPIRE1	rs8083786	0.00	0.00	0.91	0.02	0.07
STARD3	rs1877030	0.01	0.00	0.87	0.01	0.10			STARD3	rs1877030	0.01	0.00	0.89	0.02	0.08
STAT1	rs11889341	0.00	0.00	0.91	0.01	0.08			STAT1	rs11889341	0.00	0.00	0.73	0.01	0.25
STAT4	rs11889341	0.00	0.00	0.91	0.01	0.08			STAT4	rs11889341	0.00	0.00	0.90	0.01	0.09
									STRADB	rs6715284	0.07	0.00	0.84	0.02	0.08
									TAGAP	rs2451258	0.00	0.00	0.90	0.02	0.08
TCAP	rs1877030	0.01	0.00	0.89	0.01	0.08			TCAP	rs1877030	0.01	0.00	0.87	0.02	0.10
									TNFSF4	rs2105325	0.00	0.00	0.91	0.02	0.07
									TRAK2	rs6715284	0.07	0.00	0.84	0.02	0.07
TSTD1	rs4656942	0.99	0.01	0.00	0.00	0.00									
UBE2L3	rs11089637	0.11	0.00	0.80	0.00	0.08			UBE2L3	rs11089637	0.03	0.00	0.84	0.02	0.11
USF1	rs4656942	0.99	0.01	0.00	0.00	0.00									
YDJC	rs11089637	0.11	0.00	0.81	0.00	0.08			YDJC	rs11089637	0.03	0.00	0.85	0.01	0.10
YPEL1	rs11089637	0.11	0.00	0.81	0.00	0.08			YPEL1	rs11089637	0.03	0.00	0.88	0.02	0.07
ZNF335	rs4239702	0.00	0.00	0.92	0.01	0.07			ZNF335	rs4239702	0.00	0.00	0.91	0.01	0.08
ZSWIM1	rs4239702	0.00	0.00	0.91	0.01	0.08			ZSWIM1	rs4239702	0.00	0.00	0.91	0.01	0.08
ZSWIM3	rs4239702	0.00	0.00	0.92	0.01	0.07			ZSWIM3	rs4239702	0.00	0.00	0.91	0.01	0.08

Supplementary Table 6 – GSEA Enrichment Scores. [1]Enrichment Score: Degree to which gene set is overrepresented at the top of bottom of the ranked list of genes. [2]Normalized Enrichment Score: Enrichment score divided by the mean of enrichment scores against all permutations of the ranked list.

NAME	ES[1]	NES[2]	NOM p-val	FDR q-val
HALLMARK_MITOTIC_SPINDLE	-0.96953654	-2.738665	0.036809817	0.24540035
HALLMARK_TNFA_SIGNALING_VIA_NFKB	-0.78459835	-2.131841	0.061099797	1
HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSITION	-0.7627758	-2.1261334	0.064718165	0.6822187
HALLMARK_INTERFERON_GAMMA_RESPONSE	-0.75405574	-2.0704424	0.06290673	0.5373365
HALLMARK_XENOBIOTIC_METABOLISM	0.7114698	1.7093651	0.14814815	0.8671167
HALLMARK_ANGIOGENESIS	-0.7054058	-1.7550359	0.047528517	0.33538553
HALLMARK_ESTROGEN_RESPONSE_EARLY	0.70333433	1.7096115	0.14570858	1
HALLMARK_COMPLEMENT	-0.6963707	-1.8607478	0.08898305	0.5260123
HALLMARK_ESTROGEN_RESPONSE_LATE	0.69636387	1.6763769	0.13644859	0.6136178
HALLMARK_OXIDATIVE_PHOSPHORYLATION	0.68012977	1.6174641	0.17375231	0.51181173
HALLMARK_KRAS_SIGNALING_UP	-0.6740996	-1.8120219	0.09544469	0.46603143
HALLMARK_INFLAMMATORY_RESPONSE	-0.66156536	-1.8029977	0.10107527	0.4036422
HALLMARK_HYPOXIA	-0.6506489	-1.7506601	0.11594203	0.3033227

HALLMARK_INTERFERON_ALPHA_RESPONSE	-0.6433157	-1.8028818	0.057539683	0.35318694
HALLMARK_MYC_TARGETS_V1	-0.63430864	-1.6890604	0.12035011	0.30162588
HALLMARK_IL6_JAK_STAT3_SIGNALING	-0.6031751	-1.66111	0.06584362	0.28307122
HALLMARK_SPERMATOGENESIS	-0.5676277	-1.557379	0.061099797	0.28568843
HALLMARK_HEDGEHOG_SIGNALING	-0.5453301	-1.3602403	0.13793103	0.29140854
HALLMARK_APOPTOSIS	-0.5123841	-1.419991	0.111561865	0.30515093
HALLMARK_PANCREAS_BETA_CELLS	0.50091386	1.2124683	0.23517382	0.6139624
HALLMARK_GLYCOLYSIS	0.5003059	1.1435319	0.256654	0.48104674
HALLMARK_MYOGENESIS	-0.4992916	-1.3265461	0.14197531	0.27315947
HALLMARK_ALLOGRAFT_REJECTION	-0.49219608	-1.3456522	0.12778905	0.28104284
HALLMARK_UV_RESPONSE_DN	-0.4788568	-1.3921219	0.09381663	0.29563344
HALLMARK_IL2_STAT5_SIGNALING	-0.4648526	-1.2194562	0.1826087	0.32190385
HALLMARK_PEROXISOME	0.45880708	1.1639911	0.21142857	0.5433961
HALLMARK_ADIPOGENESIS	0.42392758	1.0107353	0.29338843	0.5345546
HALLMARK_G2M_CHECKPOINT	-0.42264763	-1.1181176	0.18828452	0.39563367
HALLMARK_NOTCH_SIGNALING	-0.41112712	-1.0035685	0.41614908	0.5054806

HALLMARK_TGF_BETA_SIGNALING	-0.38988957	-1.0531788	0.34435797	0.45401844
HALLMARK_BILE_ACID_METABOLISM	0.37280008	0.92135537	0.38345864	0.5846342
HALLMARK_COAGULATION	-0.35666084	-0.96529174	0.36305732	0.55080724
HALLMARK_P53_PATHWAY	-0.35356408	-0.9617617	0.28817204	0.5333001
HALLMARK_E2F_TARGETS	-0.3495585	-0.89524615	0.31905782	0.5649946
HALLMARK_FATTY_ACID_METABOLISM	0.3491053	0.86028254	0.4208494	0.6140024
HALLMARK_PI3K_AKT_MTOR_SIGNALING	-0.34672815	-0.9346718	0.4032258	0.5377159
HALLMARK_APICAL_SURFACE	-0.34615594	-0.9051891	0.53373015	0.5665438
HALLMARK_CHOLESTEROL_HOMEOSTASIS	-0.34555647	-0.9500684	0.43452382	0.5313629
HALLMARK_ANDROGEN_RESPONSE	0.31548336	0.7782141	0.5708333	0.7003621
HALLMARK_REACTIVE_OXIGEN_SPECIES_PATHWAY	-0.3002477	-0.81362534	0.6687631	0.70644766
HALLMARK_WNT_BETA_CATENIN_SIGNALING	-0.29808396	-0.766772	0.7231076	0.73594713
HALLMARK_PROTEIN_SECRETION	0.29019544	0.6923406	0.65957445	0.80639565
HALLMARK_MYC_TARGETS_V2	-0.28258362	-0.77286494	0.69753087	0.74765736
HALLMARK_DNA_REPAIR	-0.27484703	-0.77776295	0.6492693	0.7620492
HALLMARK_UV_RESPONSE_UP	-0.26478726	-0.7135497	0.73100615	0.8207212

HALLMARK_MTORC1_SIGNALING	-0.25699028	-0.6710301	0.7731481	0.87532324
HALLMARK_KRAS_SIGNALING_DN	0.25573143	0.64437884	0.83943087	0.83299124
HALLMARK_APICAL_JUNCTION	-0.23773567	-0.65685195	0.8586724	0.87192327
HALLMARK_UNFOLDED_PROTEIN_RESPONSE	0.22921126	0.56387246	0.8996139	0.90224123
HALLMARK_HEME_METABOLISM	-0.22914925	-0.6322403	0.8836689	0.88378996

Supplementary Table 7 – Ancestry Proportion Permutation Test Summary Statistics. [1]Based on the ranking of true correlation (c =0.43) out of 1000 permutation tests.

Statistic	**Value**
Minimum	-0.619
1st Quartile	-0.227
Median	0.013
Mean	-0.005
3rd Quartile	0.210
Maximum	0.567
***P* Value of True Correlation[1]**	**0.05**

Supplementary Table 8 – Cluster 16 Differentially Expressed Pathways

Upregulated				
Category	**ID**	**Name**	**p-value**	**q-value FDR B&H**
GO: Molecular Function	GO:0003823	antigen binding	2.81E-08	7.31E-06
GO: Molecular Function	GO:0034987	immunoglobulin receptor binding	8.34E-08	1.08E-05
GO: Molecular Function	GO:0051787	misfolded protein binding	6.15E-05	5.33E-03
GO: Molecular Function	GO:0038024	cargo receptor activity	1.34E-04	8.70E-03
GO: Molecular Function	GO:0060090	molecular adaptor activity	3.65E-04	1.90E-02
GO: Biological Process	GO:0042113	B cell activation	6.54E-12	1.18E-08
GO: Biological Process	GO:0002252	immune effector process	2.07E-11	1.87E-08
GO: Biological Process	GO:0050853	B cell receptor signaling pathway	8.01E-11	4.81E-08
GO: Biological Process	GO:0050864	regulation of B cell activation	1.86E-10	8.38E-08

GO: Biological Process	GO:0050871	positive regulation of B cell activation	3.49E-10	1.26E-07
GO: Biological Process	GO:0002684	positive regulation of immune system process	5.99E-10	1.80E-07
GO: Biological Process	GO:0050776	regulation of immune response	1.28E-09	3.29E-07
GO: Biological Process	GO:0051249	regulation of lymphocyte activation	2.01E-09	4.52E-07
GO: Biological Process	GO:0045321	leukocyte activation	2.61E-09	5.22E-07
GO: Biological Process	GO:0046649	lymphocyte activation	4.06E-09	6.66E-07
ToppCell Atlas	4d105437487bf04f8cd6129b822d5c060af0bee8	ILEUM-inflamed-(5)_Plasma|inflamed / shred on tissue, inflammation_status, cell class(v3), cell subclass (v2)	1.95E-77	1.28E-73
ToppCell Atlas	f3c583f8521f63834d1afbf8b07940fee7712f9a	Sigmoid-(2)_B_cell-(20)_B_cell_IgA_Plasma|Sigmoid / shred on region, Cell_type, and subtype	3.65E-76	4.80E-73
ToppCell Atlas	ea65cadaf464ff4c1040fceb5f3724706f8447c1	Transverse-B_cell-B_cell_IgA_Plasma|Transverse / Region, Cell class and subclass	3.65E-76	4.80E-73
ToppCell Atlas	634e0b614ab45e7520da67375b3498db5442b7ae	Sigmoid-B_cell-B_cell_IgA_Plasma|Sigmoid / Region, Cell class and subclass	3.65E-76	4.80E-73
ToppCell Atlas	8caaa08b046f766af298d03961ccac6e4569e2a6	Transverse-(2)_B_cell-(20)_B_cell_IgA_Plasma|Transverse / shred on region, Cell_type, and subtype	3.65E-76	4.80E-73
ToppCell Atlas	37e03f51925345f5f716ef3abe1d2c4ccf129020	tumor_Lung-B_lymphocytes-MALT_B_cells|tumor_Lung / Location, Cell class and cell subclass	1.19E-75	1.31E-72
ToppCell Atlas	a113558ca4220bac31f9336033dff82a618d9258	ILEUM-inflamed-(5)_IgA_plasma_cells|inflamed / shred on tissue, inflammation_status, cell class(v3), cell subclass (v2)	4.91E-75	4.61E-72

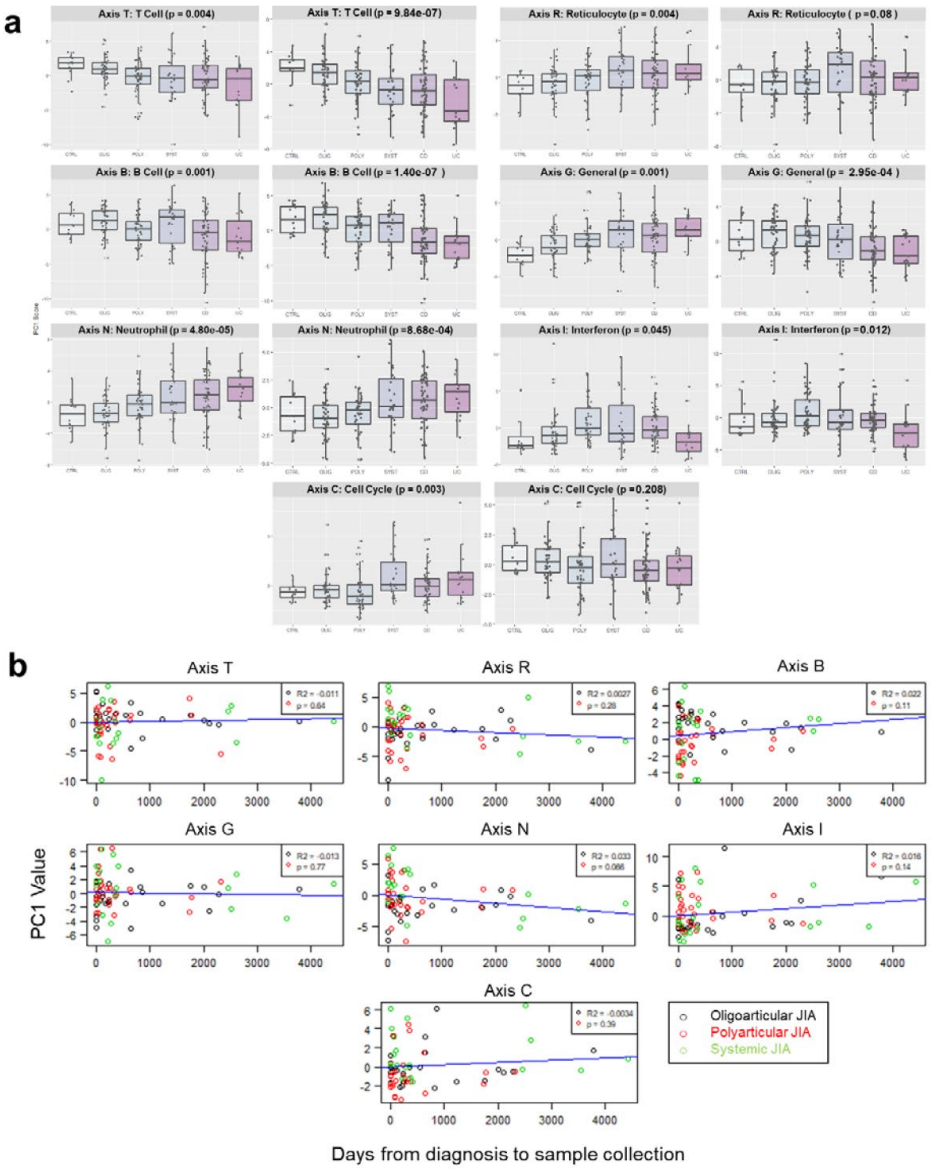
ToppCell Atlas	204e5a675f4684c9570c24c8d34bbf632fd8164a	tumor_Lymph_Node_/_Brain-B_lymphocytes-MALT_B_cells|tumor_Lymph_Node_/_Brain / Location, Cell class and cell subclass	6.76E-75	5.55E-72
ToppCell Atlas	a44e418a75fd54fd8ba266a93c0e9e3ac2e6e315	ILEUM-non-inflamed-(5)_IgA_plasma_cells|non-inflamed / shred on tissue, inflammation_status, cell class(v3), cell subclass (v2)	7.78E-75	5.68E-72
ToppCell Atlas	c85f300c76043cc10935478591c9322b78845264	ILEUM-non-inflamed-(5)_Plasma|non-inflamed / shred on tissue, inflammation_status, cell class(v3), cell subclass (v2)	1.23E-74	8.05E-72

Downregulated				
Category	**ID**	**Name**	**p-value**	**q-value FDR B&H**
GO: Molecular Function	GO:0016491	oxidoreductase activity	3.46E-45	5.34E-42
GO: Molecular Function	GO:0003954	NADH dehydrogenase activity	3.81E-38	1.47E-35
GO: Molecular Function	GO:0050136	NADH dehydrogenase (quinone) activity	3.81E-38	1.47E-35
GO: Molecular Function	GO:0008137	NADH dehydrogenase (ubiquinone) activity	3.81E-38	1.47E-35
GO: Molecular Function	GO:0016655	oxidoreductase activity, acting on NAD(P)H, quinone or similar compound as acceptor	8.23E-38	2.54E-35
GO: Molecular Function	GO:0009055	electron transfer activity	7.94E-29	2.04E-26
GO: Molecular Function	GO:0016651	oxidoreductase activity, acting on NAD(P)H	1.08E-25	2.39E-23
GO: Molecular Function	GO:0044877	protein-containing complex binding	6.59E-25	1.27E-22
GO: Molecular Function	GO:0045296	cadherin binding	4.21E-23	7.20E-21

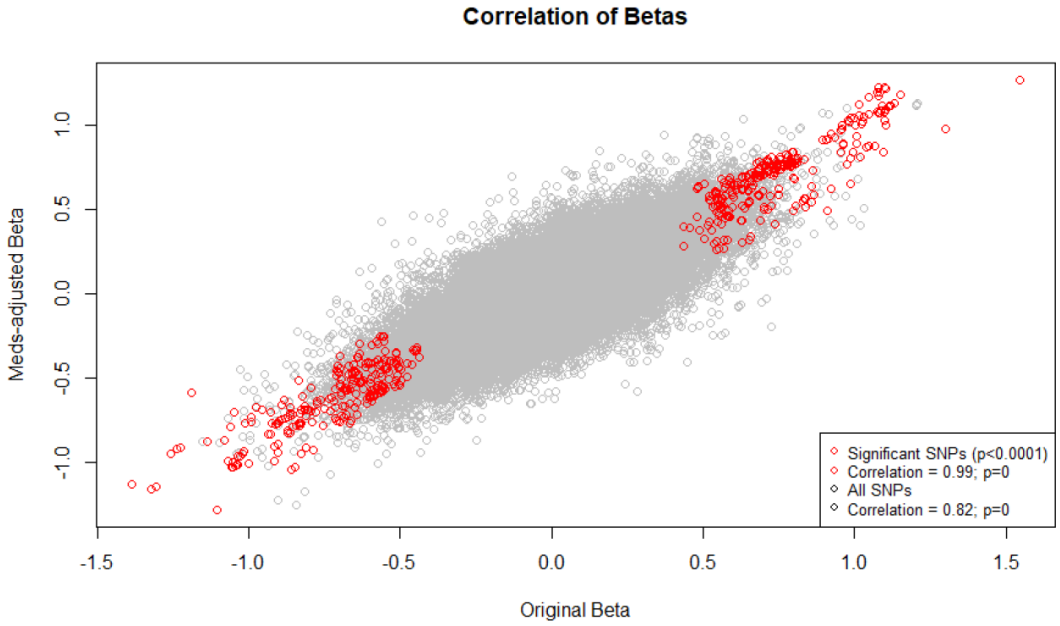
GO: Molecular Function	GO:0015078	proton transmembrane transporter activity	1.54E-22	2.37E-20
GO: Biological Process	GO:0006119	oxidative phosphorylation	1.01E-73	7.26E-70
GO: Biological Process	GO:0015986	ATP synthesis coupled proton transport	1.01E-72	2.41E-69
GO: Biological Process	GO:0015985	energy coupled proton transport, down electrochemical gradient	1.01E-72	2.41E-69
GO: Biological Process	GO:0006754	ATP biosynthetic process	4.24E-72	7.60E-69
GO: Biological Process	GO:0009206	purine ribonucleoside triphosphate biosynthetic process	7.36E-69	1.06E-65
GO: Biological Process	GO:0009145	purine nucleoside triphosphate biosynthetic process	1.33E-68	1.59E-65
GO: Biological Process	GO:0009201	ribonucleoside triphosphate biosynthetic process	2.38E-67	2.44E-64
GO: Biological Process	GO:0009142	nucleoside triphosphate biosynthetic process	3.06E-67	2.74E-64
GO: Biological Process	GO:0009144	purine nucleoside triphosphate metabolic process	4.25E-67	3.39E-64
GO: Biological Process	GO:0009205	purine ribonucleoside triphosphate metabolic process	6.28E-67	4.51E-64
GO: Cellular Component	GO:0098800	inner mitochondrial membrane protein complex	1.14E-68	1.13E-65
GO: Cellular Component	GO:0005743	mitochondrial inner membrane	1.95E-64	9.69E-62
GO: Cellular Component	GO:0005740	mitochondrial envelope	5.59E-64	1.85E-61
GO: Cellular Component	GO:0031966	mitochondrial membrane	3.18E-61	7.89E-59
GO: Cellular Component	GO:0070469	respirasome	1.37E-60	2.72E-58
GO: Cellular Component	GO:0098798	mitochondrial protein complex	2.92E-60	4.83E-58

GO: Cellular Component	GO:0031967	organelle envelope	3.6E-60	5.1E-58
GO: Cellular Component	GO:0031975	envelope	4.12E-60	5.12E-58
GO: Cellular Component	GO:0019866	organelle inner membrane	7.02E-60	7.74E-58
GO: Cellular Component	GO:0098803	respiratory chain complex	1.72E-57	1.71E-55

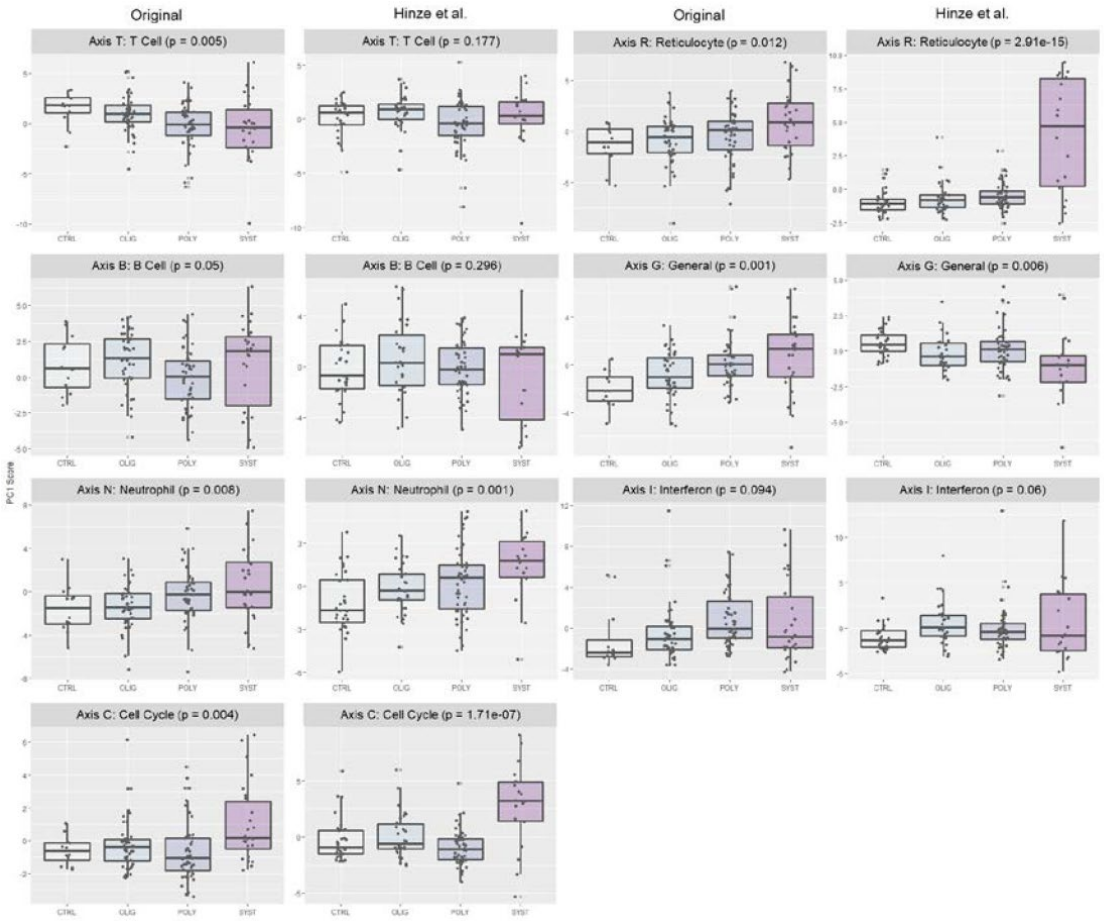
APPENDIX B. SUPPLEMENTARY FIGURES



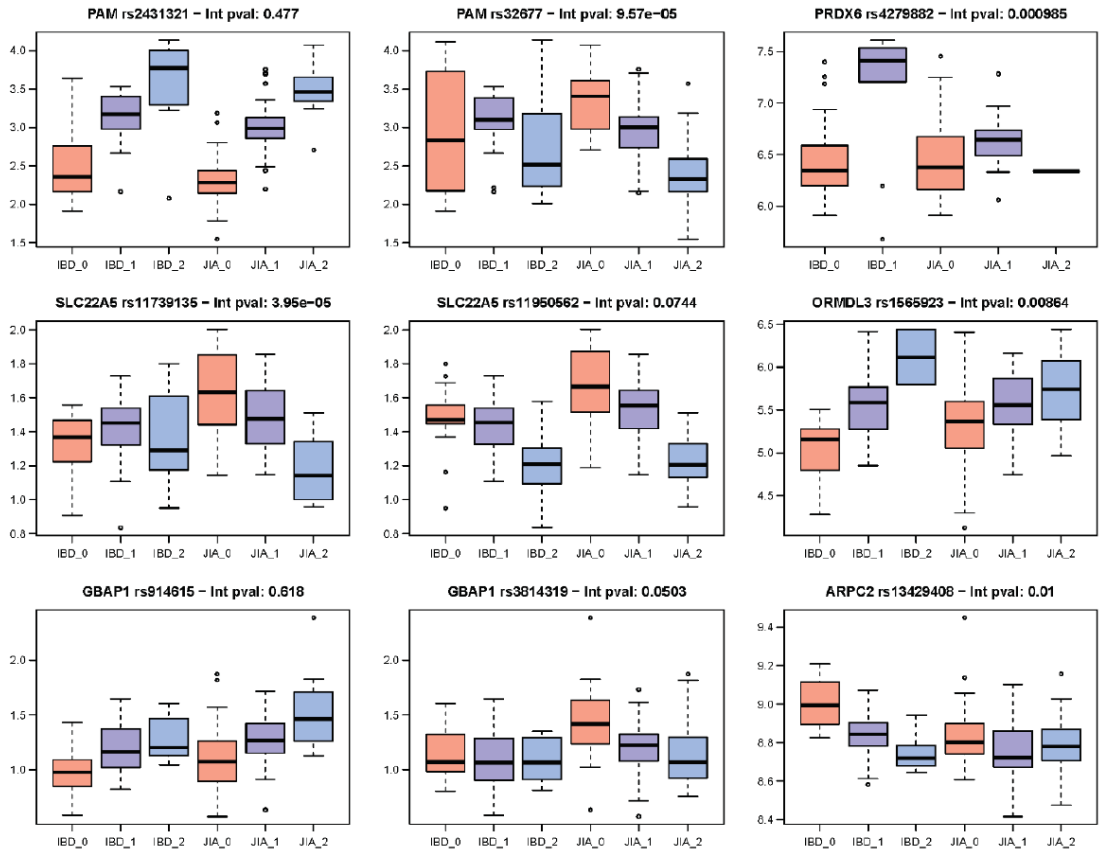
Supplementary Figure 1 – Effects of medication and sample time on gene expression. Changes in gene expression following adjustment for treatment with DMARDs, biologics, and steroids, as well as time from diagnosis to sample collection. From left to right in pairs, (a) depicts the principal component of unadjusted expression of Axis genes versus adjusted expression. (b) depicts the principal component of time to collection adjusted expression by days to sample collection.



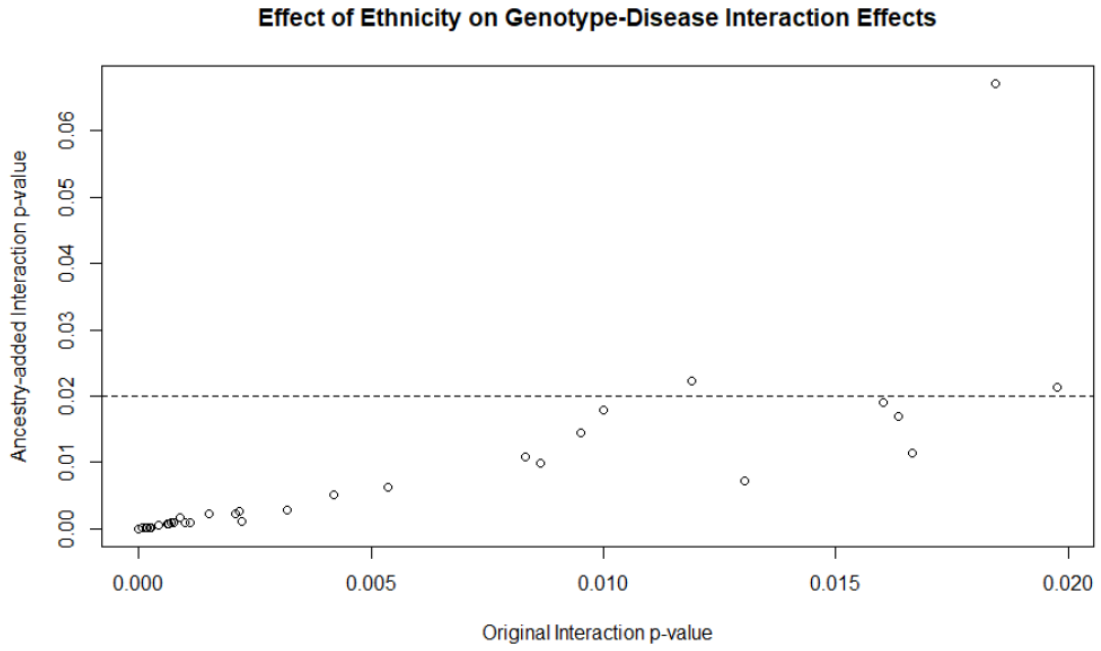
Supplementary Figure 2 – Correlation of betas in non-adjusted and medication-adjusted SNPs. Effect sizes in DMARD, biologic, and steroid treatment-adjusted SNPs versus non-adjusted SNPs. SNPs highlighted in red are significant at p < 0.0001, with a correlation = 0.99. All SNPs have a correlation of 0.82.



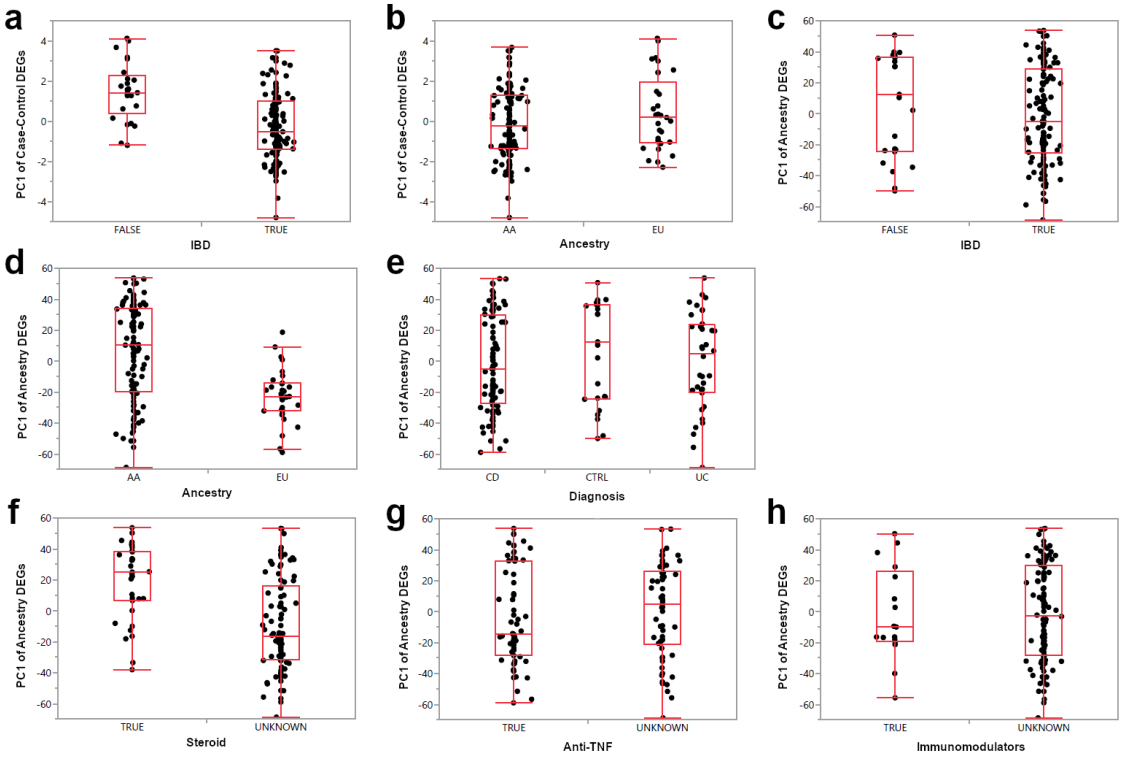
Supplementary Figure 3 – Replication of gene expression trends in the Hinze et al. dataset. The principal component of Axis gene expression in the dataset presented in this analysis, versus the principal component derived from the Hinze et al. dataset.



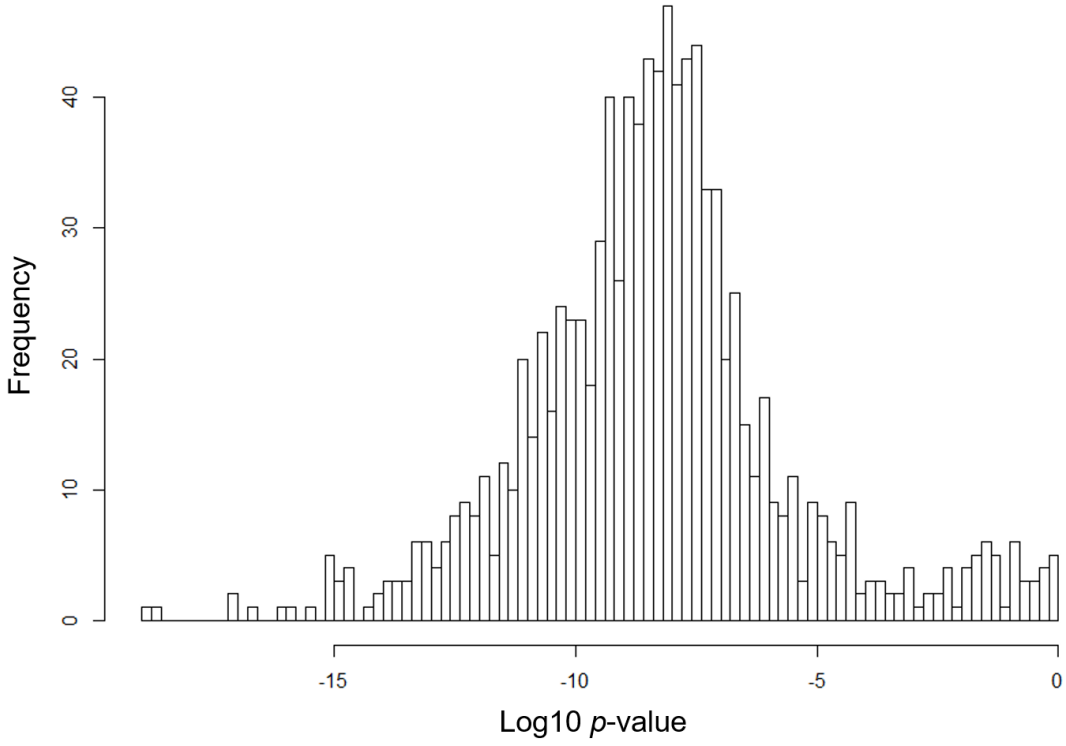
Supplementary Figure 4 – Examples of disease-specific eQTL. Selection of eQTL with JIA- or IBD-biased effects on transcript abundance. The p-values for presence of an interaction effect between genotype and disease are noted in headers.



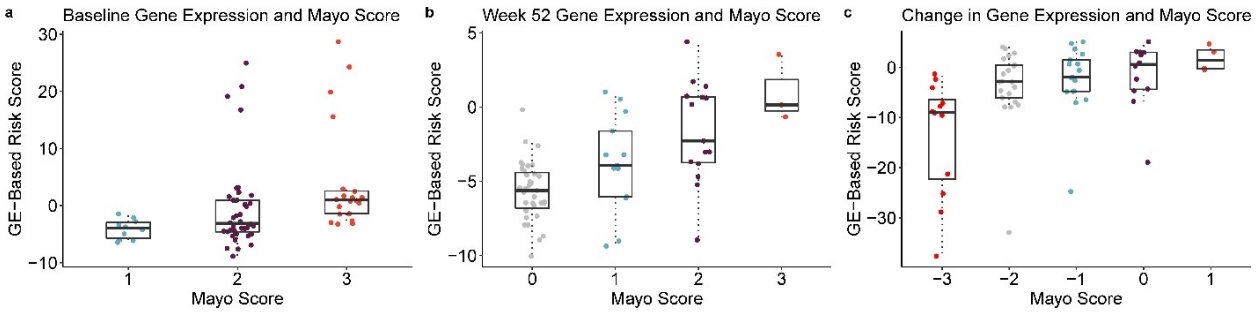
Supplementary Figure 5 – Interaction effects with addition of ethnicity. The p-value for presence of interaction effects in a model including ancestry, versus p-value in a model without. The genotype-disease interaction effects remain significant even after addition of the ethnicity term.



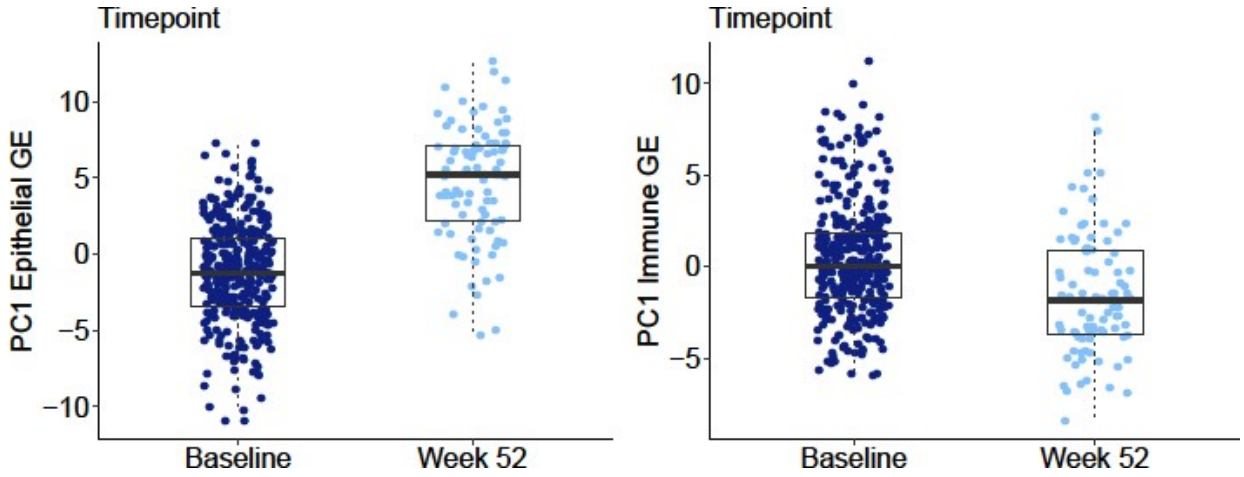
Supplementary Figure 6 – Case-control, disease subtype, and medication associations with DEGs. (a) First principal component (PC1) of genes differentially expressed between IBD cases and controls. P = 1.42×10-6 (b) PC1 of case-control DEGs in AA individuals and European individuals. P = 0.15 (c) PC1 of genes differentially expressed between AA IBD patients and European IBD patients in IBD cases and controls. P = 0.27 (d) First principal component of genes differentially expressed between AA IBD patients and European IBD patients. P = 1.38×10-6 (e) PC1 of ancestry-related DEGs in controls and subtypes of IBD. P = 0.54 (f) PC1 of ancestry-related DEGs in known patients treated with steroids. P = 9.22×10-6; note that only AA patients received steroids (g) PC1 of ancestry-related DEGs in patients known to b treated with anti-TNF. P = 0.44 (h) PC1 of ancestry-related DEGs in known patients treated with immunomodulators. P = 0.94



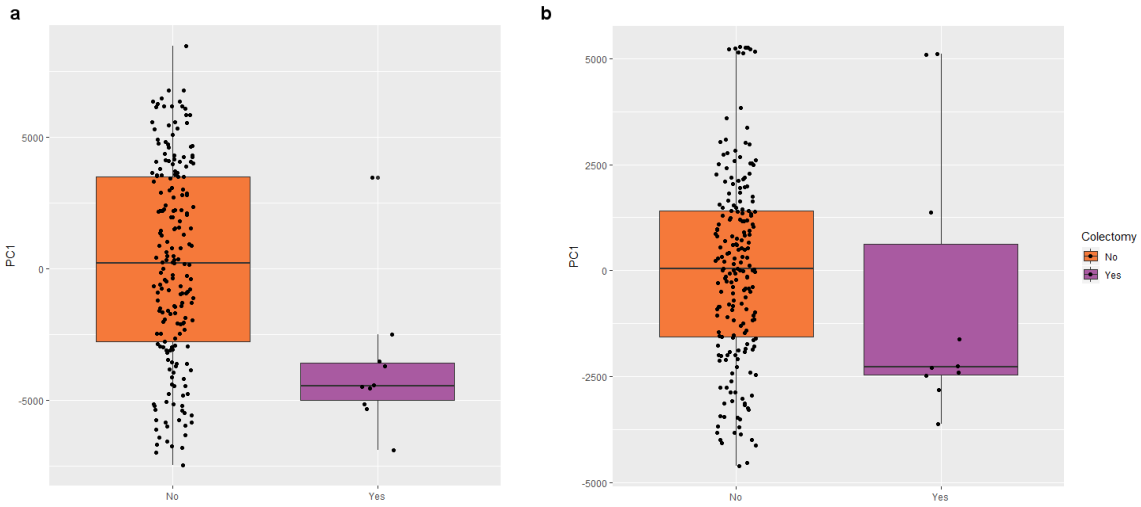
Supplementary Figure 7 – Cross-validation of $PC1_{col}$. Colectomy status was randomized prior to differential expression testing and calculation of $PC1_{colRand}$. Histogram shows frequency of log10 p-value for ANOVA test of $PC1_{colRand}$ between randomized colectomy and non-colectomy individuals in 1000 trials. $PC1_{col}$ true p = 2×10-45.



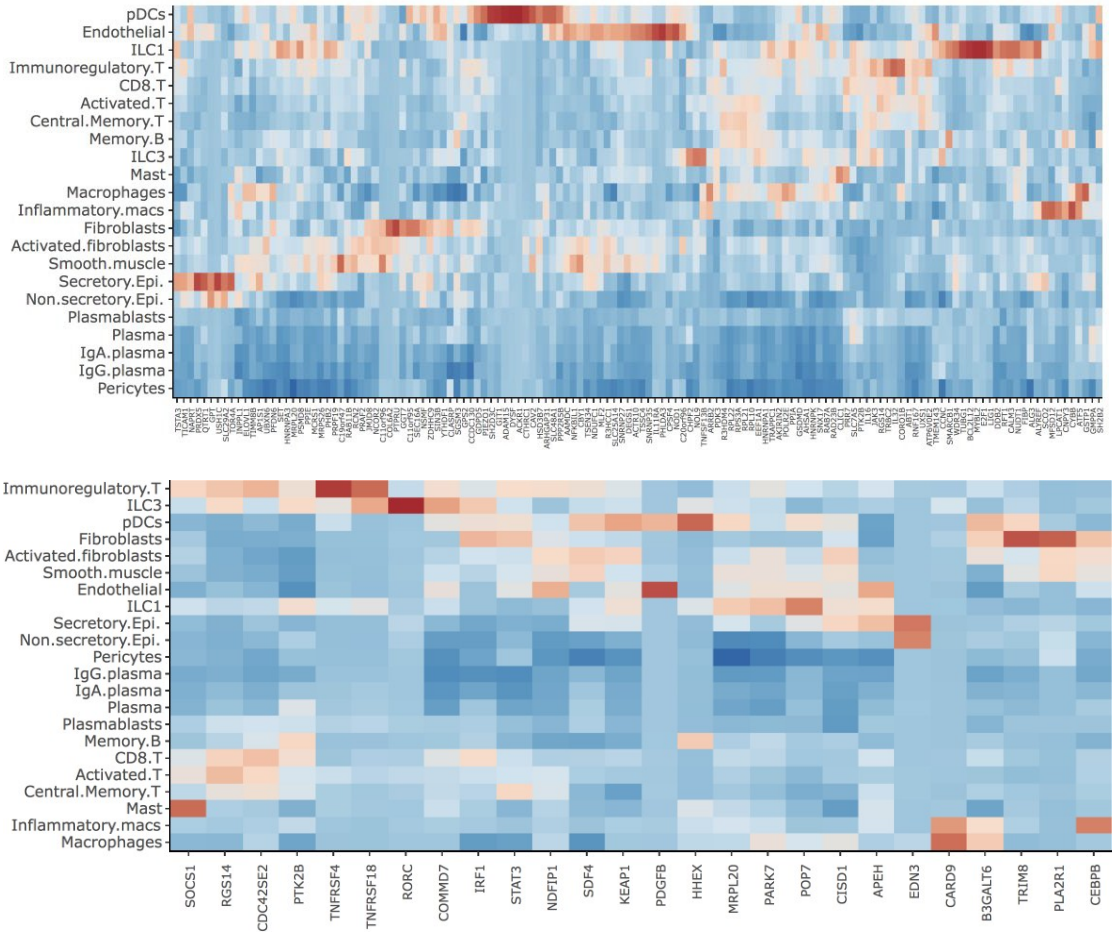
Supplementary Figure 8 – Associations between $PC1_{col}$ and Mayo score. All boxplots indicate 1st and 3rd quartile as box ends, with center median line and whiskers extending to farthest point within 1.5 times the interquartile range. (a) $PC1_{col}$ calculated on baseline gene expression with baseline Mayo score; p=0.004. (b) $PC1_{col}$ calculated on week 52 gene expression with week 52 Mayo score; p=8.73×10-8. (c) Change in $PC1_{col}$ and Mayo score from baseline to week 52; p=4×10-4.



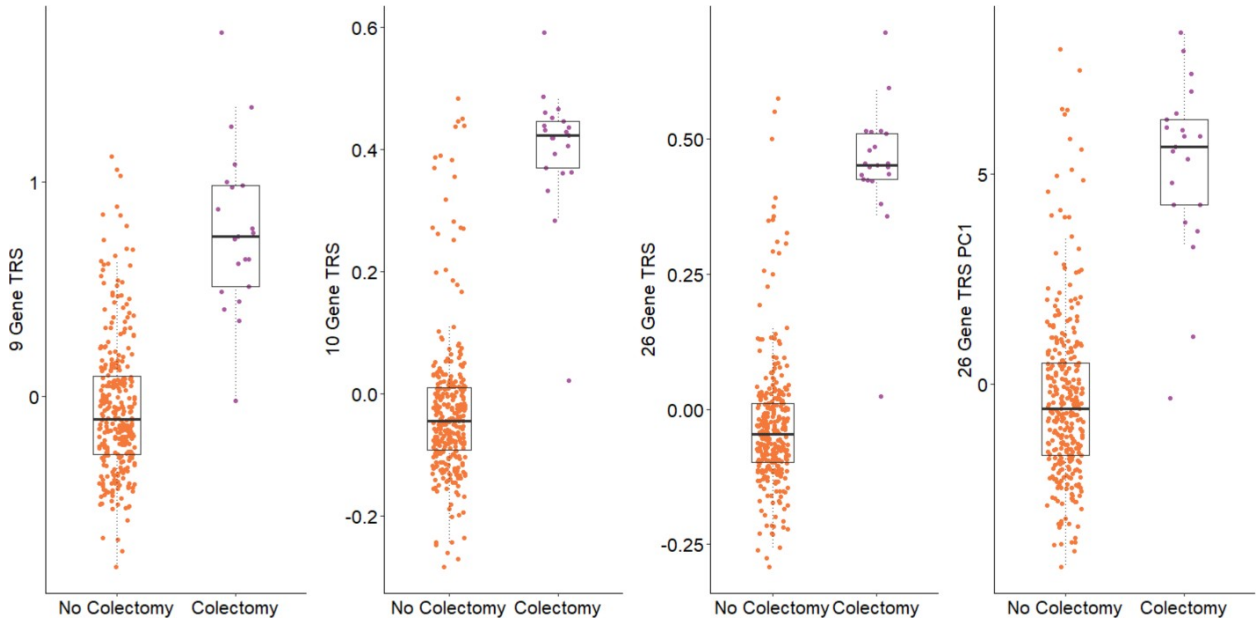
Supplementary Figure 9 – Switch in proportions of epithelial and immune components of rectal gene expression between baseline and week 52 follow-up. All boxplots indicate 1st and 3rd quartile as box ends, with center median line and whiskers extending to farthest point within 1.5 times the interquartile range. First principal components of 200 genes differentially expressed between the two tissue compartments in [Supplement ref. 27] were calculated and polarized such that PC1 reflects elevated expression of the genes. These results imply that immune activity is suppressed at week 52, and epithelial activity relatively elevated.



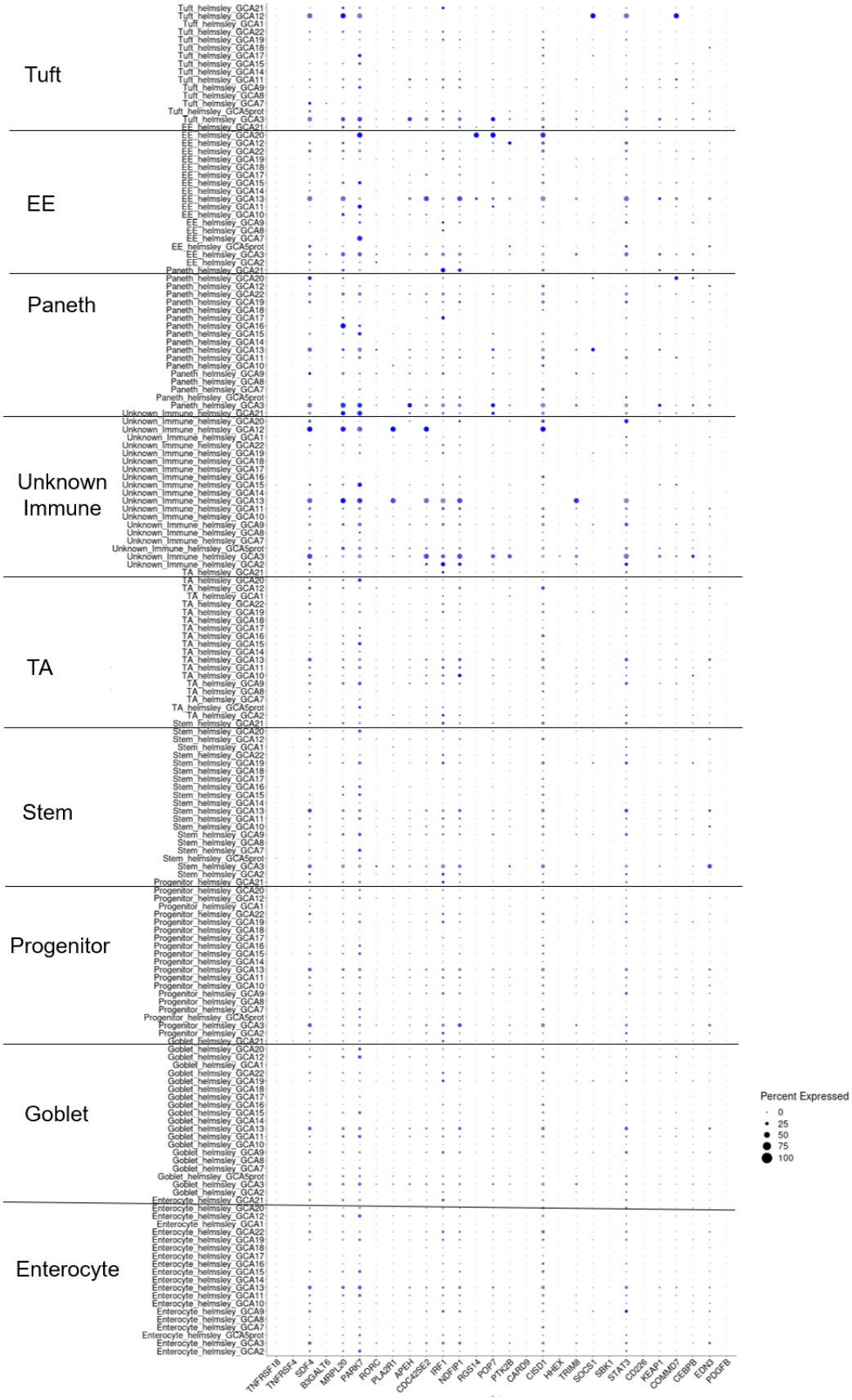
Supplementary Figure 10 – Replication of transcriptional risk prediction in the Mt Sinai cohort. All boxplots indicate 1st and 3rd quartile as box ends, with center median line and whiskers extending to farthest point within 1.5 times the interquartile range. (a) PC1 of colectomy-associated genes in Mt Sinai significantly differentiates colectomy (green) from non-colectomy (red). (b) TRSUC developed from IBD GWAS- associated genes also predicts progression to colectomy in the Mt Sinai cohort. Two outlier samples reduce the significance, which is p=0.01) for the remaining samples.



Supplementary Figure 11 – Cell-type specific expression of colectomy-associated genes. (a) Heat map showing up- regulation (red) of each gene contributing to PC1 in a rectal scRNAseq dataset. Dozens of genes are enriched in seven cell-types. (b) Similar analysis but for the TRS$_{UC}$ genes. Note the similarity of the cell- types showing enrichment, and the absence of B-cell or plasma cell signals in both.



Supplementary Figure 12 – Comparison of TRS generated with different subsets of genes. Each plot shows the computed TRS for each individual who did or did not require colectomy during the study period. All boxplots indicate 1st and 3rd quartile as box ends, with center median line and whiskers extending to farthest point within 1.5 times the interquartile range. (a) 9 gene TRS for genes significantly differentiated by status at $p<0.1$; $p=2\times10^{-25}$. (b) 10 gene TRS for genes highlighted in the text as the major clusters of up- and down-regulated in colectomy; $p=8\times10^{-43}$. (c) 26 gene TRS as sum of z-scores weighted by the magnitude of differential expression; $p=9\times10^{-49}$. (d) TRS computed simply as PC1 of the 26 genes; $p=1\times10^{-28}$.


Tuft
EE
Paneth
Unknown Immune
TA
Stem
Progenitor
Goblet
Enterocyte
TNFRSF18
TNFRSF4
SDF4
B3GALT6
MRPL20
PARK7
RORC
PLA2R1
APEH
CDC42SE2
IRF1
NDFIP1
RGS14
POP7
PTK2B
CARD9
CISD1
HHEX
TRIM8
SOCS1
SBK1
STAT3
CD226
KEAP1
COMMD7
CEBPB
EDN3
PDGFB
Percent Expressed
0
25
50
75
100

Supplementary Figure 13 – Dot plot visualizing individual-driven differences in gene expression. The x-axis consists of each of the 9 major cell type groups subdivided by individuals, while the y-axis lists TRS genes.

REFERENCES

1. Jostins L, Ripke S, Weersma RK, Duerr RH, McGovern DP, Hui KY, et al. Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. Nature. 2012;491(7422):119-24.

2. Liu JZ, van Sommeren S, Huang H, Ng SC, Alberts R, Takahashi A, et al. Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. Nat Genet. 2015;47:979.

3. Ng SC, Shi HY, Hamidi N, Underwood FE, Tang W, Benchimol EI, et al. Worldwide incidence and prevalence of inflammatory bowel disease in the 21st century: a systematic review of population-based studies. The Lancet. 2017;390(10114):2769-78.

4. Walker CH, Arora SS, Colantonio LD, Kakati DD, Fitzmorris PS, Chu DI, et al. Rates of hospitalization among African American and Caucasian American patients with Crohn's disease seen at a tertiary care center. Gastroenterology report. 2017;5(4):288-92.

5. Basson A, Swart R, Jordaan E, Mazinu M, Watermeyer G. The Association between Race and Crohn's Disease Phenotype in the Western Cape Population of South Africa, Defined by the Montreal Classification System. PLOS ONE. 2014;9(8):e104859.

6. Griglione N, Yarandi S, Srinivasan J, Ahearn T, Dhere T. A comparison of abdominal surgical outcomes between African–American and Caucasian Crohn's patients. International Journal of Colorectal Disease. 2014;29(8):917-22.

7. Kugathasan S, Denson LA, Walters TD, Kim M-O, Marigorta UM, Schirmer M, et al. Prediction of complicated disease course for children newly diagnosed with Crohn's disease: a multicentre inception cohort study. Lancet (London, England). 2017;389(10080):1710-8.

8. Marigorta UM, Denson LA, Hyams JS, Mondal K, Prince J, Walters TD, et al. Transcriptional risk scores link GWAS to eQTLs and predict complications in Crohn's disease. Nat Genet. 2017;49(10):1517-21.

9. Kaplan GG, Ng SC. Understanding and Preventing the Global Increase of Inflammatory Bowel Disease. Gastroenterology. 2017;152(2):313-21.e2.

10. Weimers P, Munkholm P. The Natural History of IBD: Lessons Learned. Current Treatment Options in Gastroenterology. 2018;16(1):101-11.

11. Staff SH. Ulcerative colitis 2013 [Available from: https://www.aboutkidshealth.ca/Article?contentid=924&language=English.

12. Turner D, Otley AR, Mack D, Hyams J, de Bruijne J, Uusoue K, et al. Development, validation, and evaluation of a pediatric ulcerative colitis activity index: a prospective multicenter study. Gastroenterology. 2007;133(2):423-32.

13. Best WR, Becktel JM, Singleton JW, Kern F, Jr. Development of a Crohn's disease activity index. National Cooperative Crohn's Disease Study. Gastroenterology. 1976;70(3):439-44.

14. Pariente B, Mary J-Y, Danese S, Chowers Y, De Cruz P, D'Haens G, et al. Development of the Lémann Index to Assess Digestive Tract Damage in Patients With Crohn's Disease. Gastroenterology. 2015;148(1):52-63.e3.

15. Baumgart DC, Sandborn WJ. Crohn's disease. The Lancet. 2012;380(9853):1590-605.

16. Rieder F, Zimmermann EM, Remzi FH, Sandborn WJ. Crohn's disease complicated by strictures: a systematic review. Gut. 2013;62(7):1072-84.

17. Farmer RG, Whelan G, Fazio VW. Long-term follow-up of patients with Crohn's disease. Relationship between the clinical pattern and prognosis. Gastroenterology. 1985;88(6):1818-25.

18. Danese S. Mechanisms of action of infliximab in inflammatory bowel disease: an anti-inflammatory multitasker. Digestive and Liver Disease. 2008;40:S225-S8.

19. Crombe V, Salleron J, Savoye G, Dupas JL, Vernier-Massouille G, Lerebours E, et al. Long-term outcome of treatment with infliximab in pediatric-onset Crohn's disease: a population-based study. Inflamm Bowel Dis. 2011;17(10):2144-52.

20. D'Haens G, Baert F, van Assche G, Caenepeel P, Vergauwe P, Tuynman H, et al. Early combined immunosuppression or conventional management in patients with newly diagnosed Crohn's disease: an open randomised trial. Lancet. 2008;371(9613):660-7.

21. Hyams J, Crandall W, Kugathasan S, Griffiths A, Olson A, Johanns J, et al. Induction and maintenance infliximab therapy for the treatment of moderate-to-severe Crohn's disease in children. Gastroenterology. 2007;132(3):863-73; quiz 1165-6.

22. Crick F. Central dogma of molecular biology. Nature. 1970;227(5258):561-3.

23. Zhao S, Fung-Leung W-P, Bittner A, Ngo K, Liu X. Comparison of RNA-Seq and Microarray in Transcriptome Profiling of Activated T Cells. PLOS ONE. 2014;9(1):e78644.

24. Oshlack A, Wakefield MJ. Transcript length bias in RNA-seq data confounds systems biology. Biology direct. 2009;4:14-.

25. Zheng W, Chung LM, Zhao H. Bias detection and correction in RNA-Sequencing data. BMC bioinformatics. 2011;12:290-.

26. Trapnell C. Defining cell types and states with single-cell genomics. Genome research. 2015;25(10):1491-8.

27. Kolodziejczyk AA, Kim JK, Svensson V, Marioni JC, Teichmann SA. The technology and biology of single-cell RNA sequencing. Mol Cell. 2015;58(4):610-20.

28. Hwang B, Lee JH, Bang D. Single-cell RNA sequencing technologies and bioinformatics pipelines. Experimental & Molecular Medicine. 2018;50(8):96.

29. Brennecke P, Anders S, Kim JK, Kołodziejczyk AA, Zhang X, Proserpio V, et al. Accounting for technical noise in single-cell RNA-seq experiments. Nature Methods. 2013;10:1093.

30. Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nature Biotechnology. 2018;36:411.

31. Trapnell C, Cacchiarelli D, Grimsby J, Pokharel P, Li S, Morse M, et al. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nature Biotechnology. 2014;32:381.

32. Westra H-J, Peters MJ, Esko T, Yaghootkar H, Schurmann C, Kettunen J, et al. Systematic identification of trans eQTLs as putative drivers of known disease associations. Nat Genet. 2013;45:1238.

33. Consortium GT, Aguet F, Brown AA, Castel SE, Davis JR, He Y, et al. Genetic effects on gene expression across human tissues. Nature. 2017;550:204.

34. Schmiedel BJ, Singh D, Madrigal A, Valdovino-Gonzalez AG, White BM, Zapardiel-Gonzalo J, et al. Impact of Genetic Polymorphisms on Human Immune Cell Gene Expression. Cell. 2018;175(6):1701-15.e16.

35. Zhernakova DV, Deelen P, Vermaat M, van Iterson M, van Galen M, Arindrarto W, et al. Identification of context-dependent expression quantitative trait loci in whole blood. Nat Genet. 2016;49:139.

36. van der Wijst MGP, Brugge H, de Vries DH, Deelen P, Swertz MA, LifeLines Cohort S, et al. Single-cell RNA sequencing identifies celltype-specific cis-eQTLs and co-expression QTLs. Nat Genet. 2018;50(4):493-7.

37. van der Wijst M, de Vries DH, Groot HE, Trynka G, Hon CC, Bonder MJ, et al. The single-cell eQTLGen consortium. Elife. 2020;9.

38. Seyhan Yazar JA-H, Kristof Wing, Anne Senabouth, M. Grace Gordon, Stacey Andersen, Qinyi Lu, Antonia Rowson, Thomas R.P. Taylor, Linda Clarke, Katia Maccora, Christine Chen, Anthony L. Cook, Chun Jimmie Ye, Kirsten A. Fairfax, Alex W. Hewitt, Joseph E. Powell. Population-scale single-cell eQTL mapping identifies cell type specific genetic control of autoimmune disease. bioRXiv. 2021.

39. de Lange KM, Moutsianas L, Lee JC, Lamb CA, Luo Y, Kennedy NA, et al. Genome-wide association study implicates immune activation of multiple integrin genes in inflammatory bowel disease. Nat Genet. 2017;49(2):256-61.

40. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, et al. Finding the missing heritability of complex diseases. Nature. 2009;461(7265):747-53.

41. Chen G-B, Lee SH, Brion M-JA, Montgomery GW, Wray NR, Radford-Smith GL, et al. Estimation and partitioning of (co)heritability of inflammatory bowel disease from GWAS and immunochip data. Hum Mol Genet. 2014;23(17):4710-20.

42. Gordon H, Trier Moller F, Andersen V, Harbord M. Heritability in inflammatory bowel disease: from the first twin study to genome-wide association studies. Inflamm Bowel Dis. 2015;21(6):1428-34.

43. Martin AR, Kanai M, Kamatani Y, Okada Y, Neale BM, Daly MJ. Clinical use of current polygenic risk scores may exacerbate health disparities. Nat Genet. 2019;51(4):584-91.

44. Chen G-B, Lee SH, Montgomery GW, Wray NR, Visscher PM, Gearry RB, et al. Performance of risk prediction for inflammatory bowel disease based on genotyping platform and genomic risk score method. BMC Med Genet. 2017;18(1):94-.

45. Hong SN, Joung J-G, Bae JS, Lee CS, Koo JS, Park SJ, et al. RNA-seq Reveals Transcriptomic Differences in Inflamed and Noninflamed Intestinal Mucosa of Crohn's Disease Patients Compared with Normal Mucosa of Healthy Controls. Inflammatory Bowel Diseases. 2017;23(7):1098-108.

46. Haberman Y, Tickle TL, Dexheimer PJ, Kim MO, Tang D, Karns R, et al. Pediatric Crohn disease patients exhibit specific ileal transcriptome and microbiome signature. J Clin Invest. 2014;124(8):3617-33.

47. Holgersen K, Kutlu B, Fox B, Serikawa K, Lord J, Hansen AK, et al. High-resolution gene expression profiling using RNA sequencing in patients with inflammatory bowel disease and in mouse models of colitis. J Crohns Colitis. 2015;9(6):492-506.

48. Corridoni D, Chapman T, Antanaviciute A, Satsangi J, Simmons A. Inflammatory Bowel Disease Through the Lens of Single-cell RNA-seq Technologies. Inflammatory Bowel Diseases. 2020;26(11):1658-68.

49. Parikh K, Antanaviciute A, Fawkner-Corbett D, Jagielowicz M, Aulicino A, Lagerholm C, et al. Colonic epithelial cell diversity in health and inflammatory bowel disease. Nature. 2019;567(7746):49-55.

50. Smillie CS, Biton M, Ordovas-Montanes J, Sullivan KM, Burgin G, Graham DB, et al. Intra- and Inter-cellular Rewiring of the Human Colon during Ulcerative Colitis. Cell. 2019;178(3):714-30.e22.

51. Stegmann A, Hansen M, Wang Y, Larsen JB, Lund LR, Ritié L, et al. Metabolome, transcriptome, and bioinformatic cis-element analyses point to HNF-4 as a central regulator of gene expression during enterocyte differentiation. Physiol Genomics. 2006;27(2):141-55.

52. Wang CC, Biben C, Robb L, Nassir F, Barnett L, Davidson NO, et al. Homeodomain factor Nkx2-3 controls regional expression of leukocyte homing coreceptor MAdCAM-1 in specialized endothelial cells of the viscera. Dev Biol. 2000;224(2):152-67.

53. Nguyen GC, Chong CA, Chong RY. National estimates of the burden of inflammatory bowel disease among racial and ethnic groups in the United States. J Crohns Colitis. 2014;8(4):288-95.

54. Wang YR, Loftus EV, Jr., Cangemi JR, Picco MF. Racial/Ethnic and regional differences in the prevalence of inflammatory bowel disease in the United States. Digestion. 2013;88(1):20-5.

55. Kurata JH, Kantor-Fish S, Frankl H, Godby P, Vadheim CM. Crohn's disease among ethnic groups in a large health maintenance organization. Gastroenterology. 1992;102(6):1940-8.

56. Deveaux PG, Kimberling J, Galandiuk S. Crohn's disease: presentation and severity compared between black patients and white patients. Dis Colon Rectum. 2005;48(7):1404-9.

57. Crandall WV, Dotson JL, Chisolm DJ, Kappelman MD. Racial Disparities in Readmission, Complications, and Procedures in Children with Crohn's Disease. Inflammatory Bowel Diseases. 2015;21(4):801-8.

58. Walker C, Allamneni C, Orr J, Yun H, Fitzmorris P, Xie F, et al. Socioeconomic Status and Race are both Independently associated with Increased Hospitalization Rate among Crohn's Disease Patients. Sci Rep. 2018;8(1):4028.

59. Samuels AD, Weese JL, Berman PM, Kirsner JB. An epidemiologic and demographic study of inflammatory bowel disease in black patients. The American Journal of Digestive Diseases. 1974;19(2):156-60.

60. Goldman CD, Kodner IJ, Fry RD, MacDermott RP. Clinical and operative experience with non-Caucasian patients with Crohn's disease. Dis Colon Rectum. 1986;29(5):317-21.

61. Tishkoff SA, Reed FA, Friedlaender FR, Ehret C, Ranciaro A, Froment A, et al. The genetic structure and history of Africans and African Americans. Science (New York, NY). 2009;324(5930):1035-44.

62. Darvasi A, Yakir B, Kuypers J, Kokoris M, Shifman S. Linkage disequilibrium patterns of the human genome across populations. Hum Mol Genet. 2003;12(7):771-6.

63. Nédélec Y, Sanz J, Baharian G, Szpiech ZA, Pacis A, Dumaine A, et al. Genetic Ancestry and Natural Selection Drive Population Differences in Immune Responses to Pathogens. Cell. 2016;167(3):657-69.e21.

64. Quach H, Rotival M, Pothlichet J, Loh Y-HE, Dannemann M, Zidane N, et al. Genetic Adaptation and Neandertal Admixture Shaped the Immune System of Human Populations. Cell. 2016;167(3):643-56.e17.

65. Gough SCL, Simmonds MJ. The HLA Region and Autoimmune Disease: Associations and Mechanisms of Action. Current genomics. 2007;8(7):453-65.

66. Beatty PG, Mori M, Milford E. Impact of racial genetic polymorphism on the probability of finding an HLA-matched donor. Transplantation. 1995;60(8):778-83.

67. Just JJ, King MC, Thomson G, Klitz W. African-American HLA class II allele and haplotype diversity. Tissue Antigens. 1996;48(6):636-44.

68. Zachary AA, Bias WB, Johnson A, Rose SM, Leffell MS. Antigen, allele, and haplotype frequencies report of the ASHI minority antigens workshops: part 1, African-Americans. Human Immunology. 2001;62(10):1127-36.

69. Loly C, Belaiche J, Louis E. Predictors of severe Crohn's disease. Scand J Gastroenterol. 2008;43(8):948-54.

70. Adler J, Rangwalla SC, Dwamena BA, Higgins PD. The prognostic power of the NOD2 genotype for complicated Crohn's disease: a meta-analysis. Am J Gastroenterol. 2011;106(4):699-712.

71. Alvarez-Lobos M, Arostegui JI, Sans M, Tassies D, Plaza S, Delgado S, et al. Crohn's disease patients carrying Nod2/CARD15 gene variants have an increased and early

need for first surgery due to stricturing disease and higher rate of surgical recurrence. Annals of surgery. 2005;242(5):693-700.

72. Meijer MJW, Mieremet-Ooms MAC, van Hogezand RA, Lamers CBHW, Hommes DW, Verspaget HW. Role of matrix metalloproteinase, tissue inhibitor of metalloproteinase and tumor necrosis factor-alpha single nucleotide gene polymorphisms in inflammatory bowel disease. World journal of gastroenterology. 2007;13(21):2960-6.

73. Cho JH. The genetics and immunopathogenesis of inflammatory bowel disease. Nature Reviews Immunology. 2008;8:458.

74. Adeyanju O, Okou DT, Huang C, Kumar A, Sauer C, Galloway C, et al. Common NOD2 Risk Variants in African Americans with Crohn's Disease Are Due Exclusively to Recent Caucasian Admixture. Inflammatory Bowel Diseases. 2012;18(12):2357-9.

75. Dassopoulos T, Nguyen GC, Talor MV, Datta LW, Isaacs KL, Lewis JD, et al. NOD2 Mutations and Anti-Saccharomyces cerevisiae Antibodies Are Risk Factors for Crohn's Disease in African Americans. The American Journal Of Gastroenterology. 2009;105:378.

76. Wang M-H, Okazaki T, Kugathasan S, Cho JH, Isaacs KL, Lewis JD, et al. Contribution of higher risk genes and European admixture to Crohn's disease in African Americans. Inflammatory Bowel Diseases. 2012;18(12):2277-87.

77. Verstockt B, Smith KG, Lee JC. Genome-wide association studies in Crohn's disease: Past, present and future. Clinical & translational immunology. 2018;7(1):e1001-e.

78. Brant SR, Okou DT, Simpson CL, Cutler DJ, Haritunians T, Bradfield JP, et al. Genome-Wide Association Study Identifies African-Specific Susceptibility Loci in African Americans With Inflammatory Bowel Disease. Gastroenterology. 2017;152(1):206-17.e2.

79. Huang C, Haritunians T, Okou DT, Cutler DJ, Zwick ME, Taylor KD, et al. Characterization of Genetic Loci That Affect Susceptibility to Inflammatory Bowel Diseases in African Americans. Gastroenterology. 2015;149(6):1575-86.

80. Somineni HK, Nagpal S, Venkateswaran S, Cutler DJ, Okou DT, Haritunians T, et al. Whole-genome sequencing of African Americans implicates differential genetic architecture in inflammatory bowel disease. The American Journal of Human Genetics. 2021;108(3):431-45.

81. Yu H, MacIsaac D, Wong JJ, Sellers ZM, Wren AA, Bensen R, et al. Market share and costs of biologic therapies for inflammatory bowel disease in the USA. Alimentary Pharmacology & Therapeutics. 2018;47(3):364-70.

82. Rutgeerts P, van Assche G, Vermeire S. Optimizing anti-TNF treatment in inflammatory bowel disease. Gastroenterology. 2004;126(6):1593-610.

83. Gibson G, Powell JE, Marigorta UM. Expression quantitative trait locus analysis for translational medicine. Genome medicine. 2015;7(1):60-.

84. Mo A, Marigorta UM, Arafat D, Chan LHK, Ponder L, Jang SR, et al. Disease-specific regulation of gene expression in a comparative analysis of juvenile idiopathic arthritis and inflammatory bowel disease. Genome medicine. 2018;10(1):48-.

85. Gutierrez-Arcelus M, Rich SS, Raychaudhuri S. Autoimmune diseases - connecting risk alleles with molecular traits of the immune system. Nature Reviews Genetics. 2016;17(3):160-74.

86. McGovern DPB, Kugathasan S, Cho JH. Genetics of Inflammatory Bowel Diseases. Gastroenterology. 2015;149(5):1163-76.e2.

87. Ye CJ, Feng T, Kwon H-K, Raj T, Wilson MT, Asinovski N, et al. Intersection of population variation and autoimmunity genetics in human T cell activation. Science. 2014;345(6202):1254665.

88. Jarvis JN, Petty HR, Tang Y, Frank MB, Tessier PA, Dozmorov I, et al. Evidence for chronic, peripheral activation of neutrophils in polyarticular juvenile rheumatoid arthritis. Arthritis Research & Therapy. 2006;8(5):R154.

89. Ogilvie EM, Khan A, Hubank M, Kellam P, Woo P. Specific gene expression profiles in systemic juvenile idiopathic arthritis. Arthritis Rheum. 2007;56(6):1954-65.

90. Barnes MG, Grom AA, Thompson SD, Griffin TA, Pavlidis P, Itert L, et al. Subtype-Specific Peripheral Blood Gene Expression Profiles in Recent-Onset Juvenile Idiopathic Arthritis. Arthritis and Rheumatism. 2009;60(7):2102-12.

91. Jiang K, Sawle AD, Frank MB, Chen Y, Wallace CA, Jarvis JN. Whole blood gene expression profiling predicts therapeutic response at six months in patients with polyarticular juvenile idiopathic arthritis. Arthritis Rheumatol. 2014;66(5):1363-71.

92. Prahalad S, Zeft AS, Pimentel R, Clifford B, McNally B, Mineau GP, et al. Quantification of the familial contribution to juvenile idiopathic arthritis. Arthritis Rheum. 2010;62(8):2525-9.

93. Ravelli A, Martini A. Juvenile idiopathic arthritis. Lancet. 2007;369(9563):767-78.

94. Macaubas C, Nguyen K, Milojevic D, Park JL, Mellins ED. Oligoarticular and polyarticular JIA: epidemiology and pathogenesis. Nature Reviews Rheumatology. 2009;5(11):616-26.

95. Mellins ED, Macaubas C, Grom AA. Pathogenesis of systemic juvenile idiopathic arthritis: some answers, more questions. Nat Rev Rheumatol. 2011;7(7):416-26.

96. Singh-Grewal D, Schneider R, Bayer N, Feldman BM. Predictors of disease course and remission in systemic juvenile idiopathic arthritis: Significance of early clinical and laboratory features. Arthritis & Rheumatism. 2006;54(5):1595-601.

97. Cui A, Quon G, Rosenberg AM, Yeung RSM, Morris Q, Consortium BS. Gene Expression Deconvolution for Uncovering Molecular Signatures in Response to Therapy in Juvenile Idiopathic Arthritis. Plos One. 2016;11(5).

98. Jarvis JN, Frank MB. Functional genomics and rheumatoid arthritis: where have we been and where should we go? Genome Med. 2010;2(7):44.

99. Wouters CH, Ceuppens JL, Stevens EA. Different circulating lymphocyte profiles in patients with different subtypes of juvenile idiopathic arthritis. Clin Exp Rheumatol. 2002;20(2):239-48.

100. Griffin TA, Barnes MG, Ilowite NT, Olson JC, Sherry DD, Gottlieb BS, et al. Gene Expression Signatures in Polyarticular Juvenile Idiopathic Arthritis Demonstrate Disease Heterogeneity and Offer a Molecular Classification of Disease Subsets. Arthritis and Rheumatism. 2009;60(7):2113-23.

101. Wong L, Jiang K, Chen Y, Hennon T, Holmes L, Wallace CA, et al. Limits of Peripheral Blood Mononuclear Cells for Gene Expression-Based Biomarkers in Juvenile Idiopathic Arthritis. Sci Rep. 2016;6:29477.

102. Barnes MG, Grom AA, Thompson SD, Griffin TA, Luyrink LK, Colbert RA, et al. Biologic Similarities Based on Age at Onset in Oligoarticular and Polyarticular Subtypes of Juvenile Idiopathic Arthritis. Arthritis and Rheumatism. 2010;62(11):3249-58.

103. Macaubas C, Nguyen K, Deshpande C, Phillips C, Peck A, Lee T, et al. Distribution of circulating cells in systemic juvenile idiopathic arthritis across disease activity states. Clin Immunol. 2010;134(2):206-16.

104. Laboratory JDRFWTDaI. Immunobase 2018 [Available from: https://www.immunobase.org.

105. Prahalad S, O'Brien E, Fraser AM, Kerber RA, Mineau GP, Pratt D, et al. Familial aggregation of juvenile idiopathic arthritis. Arthritis & Rheumatism. 2004;50(12):4022-7.

106. Hinks A, Bowes J, Cobb J, Ainsworth HC, Marion MC, Comeau ME, et al. Fine-mapping the MHC locus in juvenile idiopathic arthritis (JIA) reveals genetic heterogeneity corresponding to distinct adult inflammatory arthritic diseases. Annals of the Rheumatic Diseases. 2017;76(4).

107. Hersh AO, Prahalad S. Immunogenetics of juvenile idiopathic arthritis: A comprehensive review. Journal of Autoimmunity. 2015;64:113-24.

108. Thompson SD, Sudman M, Ramos PS, Marion MC, Ryan M, Tsoras M, et al. The susceptibility loci juvenile idiopathic arthritis shares with other autoimmune diseases extend to PTPN2, COG6, and ANGPT1. Arthritis & Rheumatism. 2010;62(11):3265-76.

109. Thompson SD, Marion MC, Sudman M, Ryan M, Tsoras M, Howard TD, et al. Genome-wide association analysis of juvenile idiopathic arthritis identifies a new susceptibility locus at chromosomal region 3q13. Arthritis & Rheumatism. 2012;64(8):2781-91.

110. Hinks A, Cobb J, Marion MC, Prahalad S, Sudman M, Bowes J, et al. Dense genotyping of immune-related disease regions identifies 14 new susceptibility loci for juvenile idiopathic arthritis. Nat Genet. 2013;45(6):664-+.

111. McIntosh LA, Marion MC, Sudman M, Comeau ME, Becker ML, Bohnsack JF, et al. Genome-Wide Association Meta-Analysis Reveals Novel Juvenile Idiopathic Arthritis Susceptibility Loci. Arthritis & Rheumatology. 2017;69(11):2222-32.

112. Stock CJW, Ogilvie EM, Samuel JM, Fife M, Lewis CM, Woo P. Comprehensive association study of genetic variants in the IL-1 gene family in systemic juvenile idiopathic arthritis. Genes & Immunity. 2008;9(4):349-57.

113. Fife MS, Gutierrez A, Ogilvie EM, Stock CJW, Samuel JM, Thomson W, et al. Novel IL10 gene family associations with systemic juvenile idiopathic arthritis. Arthritis Research & Therapy. 2006;8(5).

114. Ombrello MJ, Arthur VL, Remmers EF, Hinks A, Tachmazidou I, Grom AA, et al. Genetic architecture distinguishes systemic juvenile idiopathic arthritis from other forms of juvenile idiopathic arthritis: clinical and therapeutic implications. Annals of the Rheumatic Diseases. 2017;76(5):906.

115. Di Narzo AF, Peters LA, Argmann C, Stojmirovic A, Perrigoue J, Li K, et al. Blood and Intestine eQTLs from an Anti-TNF-Resistant Crohn's Disease Cohort Inform IBD Genetic Association Loci. Clinical and Translational Gastroenterology. 2016;7.

116. Singh T, Levine AP, Smith PJ, Smith AM, Segal AW, Barrett JC. Characterization of Expression Quantitative Trait Loci in the Human Colon. Inflammatory Bowel Diseases. 2015;21(2):251-6.

117. Kabakchiev B, Silverberg MS. Expression Quantitative Trait Loci Analysis Identifies Associations Between Genotype and Gene Expression in Human Intestine. Gastroenterology. 2013;144(7):1488-96.e3.

118. Wakil SM, Monies DM, Abouelhoda M, Al-Tassan N, Al-Dusery H, Naim EA, et al. Association of a Mutation in LACC1 With a Monogenic Form of Systemic Juvenile Idiopathic Arthritis. Arthritis & Rheumatology. 2015;67(1):288-95.

119. Assadi G, Saleh R, Hadizadeh F, Vesterlund L, Bonfiglio F, Halfvarson J, et al. LACC1 polymorphisms in inflammatory bowel disease and juvenile idiopathic arthritis. Genes and Immunity. 2016;17(4):261-4.

120. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. 2013;14(4):R36.

121. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25(16):2078-9.

122. Anders S, Pyl PT, Huber W. HTSeq-a Python framework to work with high-throughput sequencing data. Bioinformatics. 2015;31(2):166-9.

123. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26(1):139-40.

124. Leek JT, Johnson WE, Parker HS, Jaffe AE, Storey JD. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. Bioinformatics. 2012;28(6):882-3.

125. Mecham BH, Nelson PS, Storey JD. Supervised normalization of microarrays. Bioinformatics. 2010;26(10):1308-15.

126. Young MD, Wakefield MJ, Smyth GK, Oshlack A. Gene ontology analysis for RNA-seq: accounting for selection bias. Genome Biol. 2010;11(2):R14.

127. Preininger M, Arafat D, Kim J, Nath AP, Idaghdour Y, Brigham KL, et al. Blood-informative transcripts define nine common axes of peripheral blood gene expression. PLoS Genet. 2013;9(3):e1003362.

128. Li S, Rouphael N, Duraisingham S, Romero-Steiner S, Presnell S, Davis C, et al. Molecular signatures of antibody responses derived from a systems biology study of five human vaccines. Nat Immunol. 2014;15(2):195-204.

129. Inc. SI. JMP® Genomics, Version 8.0. 1989–2015.

130. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. American journal of human genetics. 2007;81(3):559-75.

131. Delaneau O, Coulonges C, Zagury J-F. Shape-IT: new rapid and accurate algorithm for haplotype inference. Bmc Bioinformatics. 2008;9.

132. Howie BN, Donnelly P, Marchini J. A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies. PLOS Genetics. 2009;5(6):e1000529.

133. Zhou X, Stephens M. Genome-wide efficient mixed-model analysis for association studies. Nat Genet. 2012;44(7):821-4.

134. Fave M-J, Lamaze FC, Soave D, Hodgkinson A, Gauvin H, Bruat V, et al. Gene-by-environment interactions in urban populations modulate risk phenotypes. Nature Communications. 2018;9.

135. Hinze CH, Fall N, Thornton S, Mo JQ, Aronow BJ, Layh-Schmitt G, et al. Immature cell populations and an erythropoiesis gene-expression signature in systemic juvenile idiopathic arthritis: implications for pathogenesis. Arthritis Research & Therapy. 2010;12(3).

136. Hu Z, Jiang K, Frank MB, Chen Y, Jarvis JN. Modeling Transcriptional Rewiring in Neutrophils Through the Course of Treated Juvenile Idiopathic Arthritis. Scientific Reports. 2018;8.

137. Giambartolomei C, Vukcevic D, Schadt EE, Franke L, Hingorani AD, Wallace C, et al. Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics. Plos Genetics. 2014;10(5).

138. Pruim RJ, Welch RP, Sanna S, Teslovich TM, Chines PS, Gliedt TP, et al. LocusZoom: regional visualization of genome-wide association scan results. Bioinformatics. 2010;26(18):2336-7.

139. Okada Y, Wu D, Trynka G, Raj T, Terao C, Ikari K, et al. Genetics of rheumatoid arthritis contributes to biology and drug discovery. Nature. 2014;506(7488):376-81.

140. Chaussabel D, Quinn C, Shen J, Patel P, Glaser C, Baldwin N, et al. A modular analysis framework for blood genomics studies: Application to systemic lupus erythematosus. Immunity. 2008;29(1):150-64.

141. Ellinghaus D, Jostins L, Spain SL, Cortes A, Bethune J, Han B, et al. Analysis of five chronic inflammatory diseases identifies 27 new associations and highlights disease-specific patterns at shared loci. Nat Genet. 2016;48(5):510-+.

142. Fairfax BP, Makino S, Radhakrishnan J, Plant K, Leslie S, Dilthey A, et al. Genetics of gene expression in primary immune cells identifies cell type-specific master regulators and roles of HLA alleles. Nat Genet. 2012;44(5):502-+.

143. Peters JE, Lyons PA, Lee JC, Richard AC, Fortune MD, Newcombe PJ, et al. Insight into Genotype-Phenotype Associations through eQTL Mapping in Multiple Cell Types in Health and Immune-Mediated Disease. PLoS Genet. 2016;12(3):e1005908.

144. Hemani G, Shakhbazov K, Westra H-J, Esko T, Henders AK, McRae AF, et al. Detection and replication of epistasis influencing transcription in humans. Nature. 2014;508(7495):249-+.

145. Mäki-Tanila A, Hill WG. Influence of gene interaction on complex trait variation with multilocus models. Genetics. 2014;198(1):355-67.

146. Newman AM, Liu CL, Green MR, Gentles AJ, Feng W, Xu Y, et al. Robust enumeration of cell subsets from tissue expression profiles. Nat Methods. 2015;12(5):453-7.

147. Huang H, Fang M, Ostins LJ, Mirkov MU, Boucher G, Anderson CA, et al. Fine-mapping inflammatory bowel disease loci to single-variant resolution. Nature. 2017;547(7662):173-+.

148. Chun S, Casparino A, Patsopoulos NA, Croteau-Chonka DC, Raby BA, De Jager PL, et al. Limited statistical evidence for shared genetic effects of eQTLs and autoimmune-disease-associated loci in three major immune-cell types. Nat Genet. 2017;49(4):600-+.

149. Tabassum R, Sivadas A, Agrawal V, Tian H, Arafat D, Gibson G. Omic personality: implications of stable transcript and methylation profiles for personalized medicine. Genome Med. 2015;7(1):88.

150. Carr EJ, Dooley J, Garcia-Perez JE, Lagou V, Lee JC, Wouters C, et al. The cellular composition of the human immune system is shaped by age and cohabitation. Nature Immunology. 2016;17(4):461-+.

151. McGonagle D, Aziz A, Dickie LJ, McDermott MF. An integrated classification of pediatric inflammatory diseases, based on the concepts of autoinflammation and the immunological disease continuum. Pediatr Res. 2009;65(5 Pt 2):38r-45r.

152. Lin YT, Wang CT, Gershwin ME, Chiang BL. The pathogenesis of oligoarticular/polyarticular vs systemic juvenile idiopathic arthritis. Autoimmun Rev. 2011;10(8):482-9.

153. Idaghdour Y, Storey JD, Jadallah SJ, Gibson G. A genome-wide gene expression signature of environmental geography in leukocytes of Moroccan amazighs. Plos Genetics. 2008;4(4).

154. Banchereau R, Hong S, Cantarel B, Baldwin N, Baisch J, Edens M, et al. Personalized Immunomonitoring Uncovers Molecular Networks that Stratify Lupus Patients (vol 165, pg 551, 2016). Cell. 2016;165(6):1548-50.

155. Mo A, Krishnakumar C, Arafat D, Dhere T, Iskandar H, Dodd A, et al. African Ancestry Proportion Influences Ileal Gene Expression in Inflammatory Bowel Disease. Cell Mol Gastroenterol Hepatol. 2020;10(1):203-5.

156. Andrews S. FastQC: a quality control tool for high throughput sequence data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc; 2010.

157. Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. Nat Methods. 2015;12(4):357-60.

158. Law CW, Chen Y, Shi W, Smyth GK. voom: precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biology. 2014;15(2):R29.

159. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences. 2005;102(43):15545-50.

160. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20(9):1297-303.

161. Auton A, Abecasis GR, Altshuler DM, Durbin RM, Abecasis GR, Bentley DR, et al. A global reference for human genetic variation. Nature. 2015;526(7571):68-74.

162. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81(3):559-75.

163. Alexander DH, Lange K. Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. BMC Bioinformatics. 2011;12(1):246.

164. Price AL, Patterson N, Hancks DC, Myers S, Reich D, Cheung VG, et al. Effects of cis and trans Genetic Ancestry on Gene Expression in African Americans. PLOS Genetics. 2008;4(12):e1000294.

165. Leijonmarck CE, Persson PG, Hellers G. Factors affecting colectomy rate in ulcerative colitis: an epidemiologic study. Gut. 1990;31(3):329-33.

166. Lee JC, Biasci D, Roberts R, Gearry RB, Mansfield JC, Ahmad T, et al. Genome-wide association study identifies distinct genetic contributions to prognosis and susceptibility in Crohn's disease. Nat Genet. 2017;49(2):262-8.

167. Hyams JS, Davis Thomas S, Gotman N, Haberman Y, Karns R, Schirmer M, et al. Clinical and biological predictors of response to standardised paediatric colitis therapy

(PROTECT): a multicentre inception cohort study. Lancet (London, England). 2019;393(10182):1708-20.

168. Haberman Y, Karns R, Dexheimer PJ, Schirmer M, Somekh J, Jurickova I, et al. Ulcerative colitis mucosal transcriptomes reveal mitochondriopathy and personalized mechanisms underlying disease severity and treatment response. Nature Communications. 2019;10(1):38.

169. Hyams JS, Davis S, Mack DR, Boyle B, Griffiths AM, LeLeiko NS, et al. Factors associated with early outcomes following standardised therapy in children with ulcerative colitis (PROTECT): a multicentre inception cohort study. The Lancet Gastroenterology & Hepatology. 2017;2(12):855-68.

170. Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nature Biotechnology. 2019;37(8):907-15.

171. Uzzan M, Martin, J., Kenigsberg E. Mapping of B cell landscape in Ulcerative Colitis lesions reveals a pathogenic response that associates with treatment resistance and disease complications. Nature Medicine. 2020 (In Submission).

172. Suárez-Fariñas M, Tokuyama M, Wei G, Huang R, Livanos A, Jha D, et al. Intestinal Inflammation Modulates the Expression of ACE2 and TMPRSS2 and Potentially Overlaps With the Pathogenesis of SARS-CoV-2-related Disease. Gastroenterology. 2021;160(1):287-301.e20.

173. Suarez-Farinas M, Huang R, Kosoy R, Irizar A, Losic B, Wei G, et al. Tu1802 - Disease Demarcation in Ulcerative Colitis is Associated with Different Patterns of Gene Expression. Gastroenterology. 2018;154:S-1024.

174. Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WM, 3rd, et al. Comprehensive Integration of Single-Cell Data. Cell. 2019;177(7):1888-902.e21.

175. Martin JC, Boschetti G, Chang C, Ungaro R, Giri M, Chuang L-S, et al. Single-cell analysis of Crohn's disease lesions identifies a pathogenic cellular module associated with resistance to anti TNF therapy. bioRxiv. 2018:503102.

176. Wickham H. ggplot2: Elegant Graphics for Data Analysis: Springer-Verlag New York; 2016. Available from: https://ggplot2.tidyverse.org.

177. Graham DB, Xavier RJ. Pathway paradigms revealed from the genetics of inflammatory bowel disease. Nature. 2020;578(7796):527-39.

178. Turner D, Hyams J, Markowitz J, Lerer T, Mack DR, Evans J, et al. Appraisal of the pediatric ulcerative colitis activity index (PUCAI). Inflamm Bowel Dis. 2009;15(8):1218-23.

179. Peters LA, Perrigoue J, Mortha A, Iuga A, Song W-m, Neiman EM, et al. A functional genomics predictive network model identifies regulators of inflammatory bowel disease. Nat Genet. 2017;49(10):1437-49.

180. Mowat AM, Agace WW. Regional specialization within the intestinal immune system. Nat Rev Immunol. 2014;14(10):667-85.

181. Tang F, Barbacioru C, Wang Y, Nordman E, Lee C, Xu N, et al. mRNA-Seq whole-transcriptome analysis of a single cell. Nature Methods. 2009;6(5):377-82.

182. Bigaeva E, Uniken Venema WTC, Weersma RK, Festen EAM. Understanding human gut diseases at single-cell resolution. Hum Mol Genet. 2020;29(R1):R51-R8.

183. Haber AL, Biton M, Rogel N, Herbst RH, Shekhar K, Smillie C, et al. A single-cell survey of the small intestinal epithelium. Nature. 2017;551(7680):333-9.

184. Huang B, Chen Z, Geng L, Wang J, Liang H, Cao Y, et al. Mucosal Profiling of Pediatric-Onset Colitis and IBD Reveals Common Pathogenics and Therapeutic Pathways. Cell. 2019;179(5):1160-76.e24.

185. Uniken Venema WT, Voskuil MD, Vila AV, van der Vries G, Jansen BH, Jabri B, et al. Single-Cell RNA Sequencing of Blood and Ileal T Cells From Patients With Crohn's Disease Reveals Tissue-Specific Characteristics and Drug Targets. Gastroenterology. 2019;156(3):812-5.e22.

186. Kinchen J, Chen HH, Parikh K, Antanaviciute A, Jagielowicz M, Fawkner-Corbett D, et al. Structural Remodeling of the Human Colonic Mesenchyme in Inflammatory Bowel Disease. Cell. 2018;175(2):372-86.e17.

187. Wang Y, Song W, Wang J, Wang T, Xiong X, Qi Z, et al. Single-cell transcriptome analysis reveals differential nutrient absorption functions in human intestine. J Exp Med. 2020;217(2).

188. Elmentaite R, Ross ADB, Roberts K, James KR, Ortmann D, Gomes T, et al. Single-Cell Sequencing of Developing Human Gut Reveals Transcriptional Links to Childhood Crohn’s Disease. Developmental Cell. 2020;55(6):771-83.e5.

189. Peterson LW, Artis D. Intestinal epithelial cells: regulators of barrier function and immune homeostasis. Nature Reviews Immunology. 2014;14(3):141-53.

190. Gersemann M, Becker S, Kübler I, Koslowski M, Wang G, Herrlinger KR, et al. Differences in goblet cell differentiation between Crohn's disease and ulcerative colitis. Differentiation. 2009;77(1):84-94.

191. Zheng GXY, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Wilson R, et al. Massively parallel digital transcriptional profiling of single cells. Nature Communications. 2017;8(1):14049.

192. Chen J, Bardes EE, Aronow BJ, Jegga AG. ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. Nucleic Acids Res. 2009;37(Web Server issue):W305-W11.

193. Grün D, Lyubimova A, Kester L, Wiebrands K, Basak O, Sasaki N, et al. Single-cell messenger RNA sequencing reveals rare intestinal cell types. Nature. 2015;525(7568):251-5.

194. Kim YS, Ho SB. Intestinal goblet cells and mucins in health and disease: recent insights and progress. Curr Gastroenterol Rep. 2010;12(5):319-30.

195. Dorofeyev AE, Vasilenko IV, Rassokhina OA, Kondratiuk RB. Mucosal Barrier in Ulcerative Colitis and Crohn's Disease. Gastroenterology Research and Practice. 2013;2013:431231.

196. Mattei J, Parnell LD, Lai CQ, Garcia-Bailo B, Adiconis X, Shen J, et al. Disparities in allele frequencies and population differentiation for 101 disease-associated single nucleotide polymorphisms between Puerto Ricans and non-Hispanic whites. BMC Genet. 2009;10:45.

197. Williams AL, Jacobs SBR, Moreno-Macías H, Huerta-Chagoya A, Churchhouse C, Márquez-Luna C, et al. Sequence variants in SLC16A11 are a common risk factor for type 2 diabetes in Mexico. Nature. 2014;506(7486):97-101.

198. Haiman CA, Chen GK, Blot WJ, Strom SS, Berndt SI, Kittles RA, et al. Genome-wide association study of prostate cancer in men of African ancestry identifies a susceptibility locus at 17q21. Nat Genet. 2011;43(6):570-3.

199. Duncan L, Shen H, Gelaye B, Meijsen J, Ressler K, Feldman M, et al. Analysis of polygenic risk score usage and performance in diverse human populations. Nature Communications. 2019;10(1):3328.